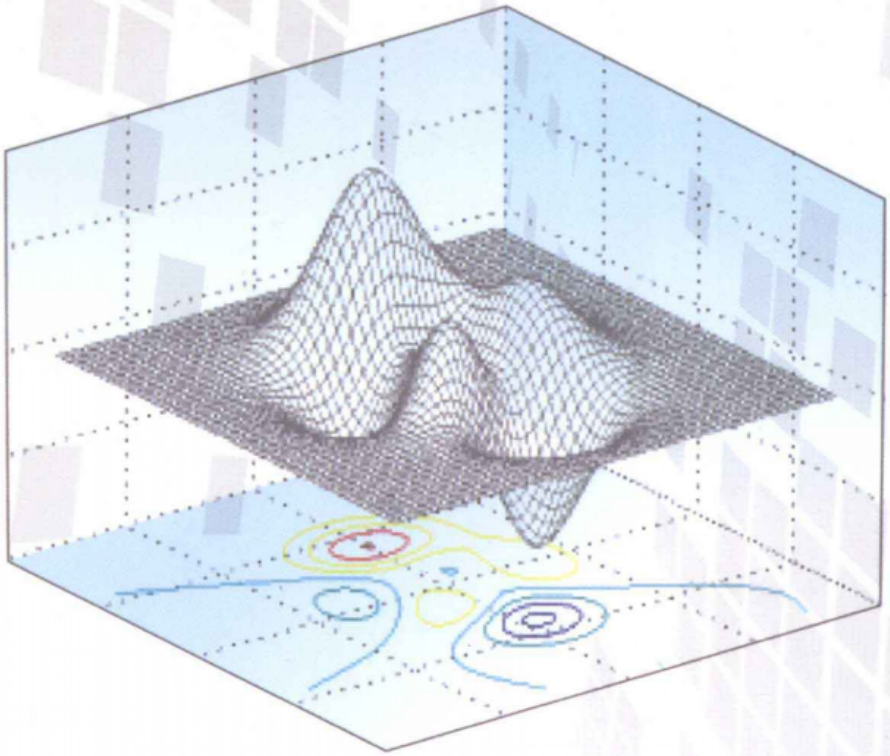
中国气象局培训中心培训系列教材

A Textbook on Support Vector Machines

支持向量机方法

应用教程

陈永义 熊秋芬 编著



气象出版社
China Meteorological Press

支持向量机方法应用教程

ISBN 978-7-5029-5236



定价：36.00元

中国气象局培训中心培训系列教材

支持向量机方法应用教程

陈永义　熊秋芬　编著

气象出版社
China Meteorological Press

内容简介

支持向量机(Support Vector Machines,SVM)是令人瞩目的机器学习新方法,但其严密的理论结构又让应用者望而却步。本书是在作者多年进行的SVM应用方法研究、教学和推广的基础上完成的一本实用教程。

本书以SVM的应用为主线,用通俗、直观的语言,由浅入深地介绍了SVM分类和回归的基础原理、实现方法及特点,对学习机性能评估和统计学习理论也有所涉猎。书中提供了一个应用SVM方法学习建模和预报应用的软件平台,使读者能够快速"上手"为我所用地解决实际问题,并给出了天气预报中的应用实例。

本书对在广泛领域中,涉及分类、回归等问题的工程技术人员和研究工作者有参考价值,也可作为计算机、信息等相关专业本科生和研究生的SVM选修课教材。

图书在版编目(CIP)数据

支持向量机方法应用教程/陈永义,熊秋芬编著.—北京:气象出版社,2011.7

ISBN 978-7-5029-5236-5

Ⅰ.①支… Ⅱ.①陈… ②熊… Ⅲ.①向量计算机—算法理论—教材 Ⅳ.①TP301.6

中国版本图书馆CIP数据核字(2011)第114690号

支持向量机方法应用教程

Zhichixiangliangji Fangfa Yingyong Jiaocheng

出版发行:气象出版社

地　　址:北京市海淀区中关村南大街46号　　**邮政编码**:100081

总 编 室:010-68407112　　**发 行 部**:010-68409198

网　　址:http://www.cmp.cma.gov.cn　　**E-mail**:qxcbs@cma.gov.cn

责任编辑:何晓欢　张　斌　　**终　　审**:黄润恒

封面设计:博雅思企划　　**责任技编**:吴庭芳

印　　刷:北京京科印刷有限公司

开　　本:720 mm×960 mm　1/16　　**印　　张**:10.5

字　　数:215千字　　**印　　张**:1～1 500

版　　次:2011年7月第1版　　**印　　次**:2011年7月第1次印刷

定　　价:36.00元

本书如存在文字不清、漏印以及缺页、倒页、脱页等,请与本社发行部联系调换

前　言

“计算机学习”是当前国际上计算机理论和应用研究的最热门话题之一。尽管至今仍有学者质疑计算机能否像人一样“学习”，但是让计算机基于已有数据建立分类、回归等模型，从而应用于实际预报、预测、诊断、分析和决策等的研究，近年来已有长足进展。而支持向量机方法(Support Vector Machines，SVM)则是其中最受瞩目的一种新方法。

2000 年，我初次接触 SVM 方法，立即被它的新颖思想所吸引——把样本空间“升维”，在无穷维空间中来解决问题，却又不增加计算的复杂性。然而，使这一切得以实现的关键竟是依据近百年前的一个纯数学定理。新的统计学习理论和 SVM 方法不愧是使机器学习研究腾飞的强劲双翅——前者是严密的理论结构，奠定了机器学习的理论基石，而后者是理论的巧妙实现，给出了有效的应用新方法。两翼双飞，二者结合得如此完美。我相信这一新方法能够在广阔的应用领域大显身手，自然也会在天气预报，特别是数值预报产品的释用上得到成功的应用。

2002 年成都市气象台的冯汉中同志来中国气象局培训中心访问进修，他很快地掌握了 SVM 方法并用之成功实验了降水分类预报和单站气温预报，进而用 VB 语言开发出了基于 SVM^{Light} 的 SVM 学习建模、预报的应用软件平台 CMSVM 的雏形。本书的初稿就是在这一过程中形成的培训讲义(2003 年)。

随后，中国气象局培训中心将 SVM 方法及其在气象中的应用作为重要的新技术之一，先后安排在省级首席预报员培训班、高级工程师培训班、数值预报释用班、气候系统监测及短期气候预测培训班、城市环境班、气象普及班、民航气象班、研究生班及 SVM 方法专题班等系列培训班进行讲授和推广，百余个气象台站和气象科研院所正式将 SVM 方法和 CMSVM 软件引入业务和科研中应用，国家气象中心的刘还珠老师还把 SVM 方法嵌入国家气象中心的“气象要素客观集成系统”加以推广应用。本书附录列出的论文索引从一个侧面反映了 SVM 方法近年来在气象业务中的推广应用情况。

为了便于推广应用，我们把相关培训教材和应用软件 CMSVM 1.0

公开放在了网站上(http://streaml.cma.gov.cn/cmsvm/)。据我们收到的可靠反馈信息的不完全统计,它们还得到了计算机、遥感、水文、地质、地震、海洋、烟草、电力、农业、医学、经济和金融等领域的专业人员的关注和使用,并有南京信息工程大学、解放军理工大学、成都信息工程学院、云南大学、中山大学、中南林业科技大学、吉林大学、哈尔滨工业大学、华东师范大学、上海海事大学、西安电子科技大学、西安科技大学、北京科技大学、南京大学、中国科学院大气物理研究所等不同专业院校所的研究生应用我们的 CMSVM 做研究。

目前提供的这本书是我们多年的读书学习、应用研究和培训教学工作的一个粗略总结。SVM 方法涉及较多的数学基础知识和理论,如泛函分析、统计分析、最优化理论、最优化算法等。作为面向气象及其他应用领域研究人员的一本应用教材,本书内容重点放在了讲清 SVM 的基本思想和具体应用方法上;而未能对相关的理论问题(如理论证明、算法实现、模型泛化能力界的估计依据等)加以全面介绍和深入讨论,是本书的缺憾之一。值得庆幸的是本书所引用的主要外文参考文献近年已相继有中译本出版,这些书是有志于从事核方法和机器学习理论研究的人员的好参考,有关内容都可以从中找到答案。

本书第 1～6 章力图对 Vapnik V N 四十年研究的成果,给出尽可能通俗、直观的介绍;第 7 章介绍了我们开发的一个应用 SVM 方法做学习建模、预报的软件平台 CMSVM 2.0,该软件可在网上(http://streaml.cma.gov.cn/cmsvm)免费下载;第 8 章介绍了 SVM 方法在气象预报中的初步应用;附录给出了截至 2009 年年底公开发表的有关 SVM 方法在我国气象领域应用文章的目录索引。

本书虽然主要是面向气象的培训教程,但所介绍的理论和方法具有一定的通用性。天气预报常遇到的分类、回归等问题,在水文、海洋、地质、地震、医学、农业和人文等广泛的领域同样会遇到。“辞异而理同”,本书介绍的方法、实例和软件工具 CMSVM 2.0,对多领域的工程技术人员和研究工作者均具有参考价值。

本书也可作为计算机、信息等相关专业本科高年级和研究生 SVM 选修课教材之用。如果重点选讲前 6 章,辅以 CMSVM 软件系统的应用实践,大约需要 40 学时。

本书第 1～7 章为陈永义撰写,第 8 章由熊秋芬执笔。

本书作者感谢国家自然科学基金(60072006)的资助，感谢中国气象局培训中心高学浩主任和俞小鼎教授的鼓励、关注和支持。向所有被引用的参考文献的作者一并表示诚挚谢意。

还要感谢王泳、李春辉在CMSVM软件平台实现上的辛勤劳动，由于他们的出色工作，使CMSVM 2.0软件得以顺利在气象系统推广应用。

限于作者们的水平，本书的内容难免会有疏漏，殷切地期望读者对本书提出修改意见。

陈永义
2010年7月于北京
yongyichen45@gmail.com

本书阅读建议

本书内容的层次结构安排，考虑到了不同读者的需求。

如果读者想对SVM方法及其应用做一粗略、宏观的了解并不实际应用，可以只阅读第1章。它对SVM方法的基本思想、特点及应用做了通俗概括的介绍。

如果读者熟悉并应用过其他的机器学习或统计分析方法（如判别分析、回归分析等），则可以在阅读第1章后直接进入第7章。第7章是我们开发的一个应用SVM方法的软件平台。熟悉CMSVM 2.0软件平台后，可直接用SVM方法解决实际应用问题并加以比较，然后根据感受再决定是否阅读其他章节。

第1～4章介绍了SVM方法的核心知识。如果读者想掌握该方法并应用其解决实际应用问题，就必须仔细阅读这4章，并要把它们基本读懂。在此基础上直接进入第7章，熟悉CMSVM软件平台后，就可以试用SVM来解决实际应用问题。

读者初步应用SVM方法解决分类或回归问题后，可能会产生一系列新问题。第5章的内容就是针对此情况撰写的，它帮助读者在几个方面对SVM方法作进一步的思考，逐步深入到计算机学习理论的核心问题。

第6章概略介绍了计算机学习的理论问题，是本书最难读的一章，对于只关心实际应用的读者可以跳过它。但对于关心统计学习理论问题的读者来说，这一章还算是较“通俗”的介绍。由于介绍得太过粗略，要彻底搞清楚这些理论问题，还必须补充阅读其他的书籍（如Vapnik的著作）。

第8章给出了几个天气预报中的应用实例和对不同方法的比较分析。虽然内容只限于气象领域，但触类旁通，对其他领域的应用也有参考价值。

当然，读者如果有兴趣并具备条件，最好的阅读建议是有重点地循序渐进地通读全书。这可以让你不但对SVM方法有较全面的了解，并能很好掌握一种软件工具，灵活方便地解决有关实际问题。

目　录

前言

本书阅读建议

第 1 章　绪　论 …… (1)

1.1　计算机应用的历史回顾 …… (1)

1.2　计算机学习的基本问题 …… (2)

1.3　SVM 方法的基本思想 …… (6)

1.4　SVM 方法的特点和应用展望 …… (8)

1.5　SVM 方法的参数优化 …… (10)

1.6　本章小结 …… (10)

第 2 章　线性支持向量机模式识别 …… (11)

2.1　模式识别问题的表述 …… (11)

2.2　最优划分超平面与支持向量的概念 …… (13)

2.3　最优划分超平面的求解 …… (15)

2.4　线性不可分问题的求解 …… (20)

2.5　线性多类分类问题的求解 …… (22)

2.6　本章小结 …… (26)

第 3 章　非线性支持向量机模式识别 …… (27)

3.1　数学预备知识 …… (28)

3.2　非线性映射与特征空间 …… (32)

3.3　特征空间中的线性学习机 …… (34)

3.4　Mercer 核和内积 …… (35)

3.5　基于核方法的非线性 SVM …… (38)

3.6　SVM 方法的特点 …… (39)

3.7　本章小结 …… (41)

第 4 章　支持向量机回归分析 …… (42)

4.1　回归分析的问题表述 …… (42)

4.2　ε-不敏感函数 …… (43)

4.3　最优回归超平面与 SVM 线性回归 …… (44)

4.4　非线性 SVM 回归 …… (48)

4.5　SVM 回归方法的特点 …… (50)

4.6　本章小结 …… (54)

第 5 章 关于支持向量机方法的进一步思考 …… (55)
5.1 从样本到样本的推理 …… (55)
5.2 SVM 方法的非线性本质 …… (56)
5.3 关于核方法 …… (57)
5.4 学习机性能的评价 …… (60)
5.5 标准 SVM 及其变种 …… (62)
5.6 SVM 方法的弱点和“开问题” …… (63)
5.7 本章小结 …… (63)
第 6 章 统计学习理论 …… (65)
6.1 ERM 归纳原则 …… (65)
6.2 统计学习的基本定理 …… (67)
6.3 函数集的 VC 维 …… (70)
6.4 VC 维与生长函数的关系 …… (76)
6.5 关于学习机推广能力的界 …… (77)
6.6 SRM 归纳原则 …… (81)
6.7 本章小结 …… (83)
第 7 章 CMSVM 2.0 软件平台及其使用方法 …… (84)
7.1 CMSVM 软件平台的设计思路 …… (84)
7.2 CMSVM 2.0 软件平台概述 …… (87)
7.3 CMSVM 2.0 的安装及系统目录结构 …… (88)
7.4 数据文件格式要求 …… (89)
7.5 CMSVM 2.0 使用说明 …… (92)
7.6 运行生成文件和对应查询 …… (109)
7.7 关于 CMSVM 系统中的核函数 …… (118)
7.8 关于 CMSVM 系统中的其他参数 …… (122)
7.9 关于 CMSVM 系统中的参数寻优 …… (124)
7.10 本章小结 …… (128)
第 8 章 支持向量机方法在天气预报中的应用 …… (129)
8.1 SVM 方法用于降水预报和温度预报的建模实例 …… (130)
8.2 SVM 方法用于预报因子筛选的实例 …… (135)
8.3 SVM 方法用于短期气候预测 …… (146)
8.4 SVM 与 ANN 方法预报效果的比较 …… (148)
8.5 本章小结 …… (151)
参考文献 …… (152)
附录 支持向量机方法在天气预报中应用论文索引 …… (154)

第1章 绪 论

1.1 计算机应用的历史回顾

在计算机[①]诞生以来的五十多年里，它给人类社会和人们生活带来了难以言尽的巨大变化。从应用角度来看，这半个世纪计算机技术的发展，大体上可以划分为三个阶段，或者说是历经了三个台阶——数值计算、数据处理和知识处理。

数值计算是人们发明计算机的初衷，希望借助它使人们能计算规模巨大的数值问题，并且算得更快、更准确。在这一方面人们已经取得了极大的成功，甚至可以说在数值计算的主要指标上，“电脑”已经超过了人脑。

如果计算机的应用只限于数值计算，那它不过是一个“大算盘”。人们很快就发现计算机还善于进行比数值计算更广义的数据处理。不但数值是数据，符号、文字、声音、图像等都是数据，计算机可以对这类广义的数据进行大容量、高速、高效、精确的管理。随着数据库理论的成熟和广泛应用，大量的数据库系统(Data Base System，DBS)、管理信息系统(Management Information System，MIS)、企业资源计划系统(Enterprise Resource Planning，ERP)、办公自动化与网络系统等几乎渗入各行各业，进入人类活动的各个领域，取得了难以估量的社会效益和经济效益。人们生产生活的各个方面都离不开计算机的数据处理。计算机在数据处理上取得了巨大的成功。

人类绝不满足于已有的成功。透过成功的眩目光环，人们发现几乎全部前面提到的成功应用，都仅局限于对数据的简单处理上，无非是对原始数据的录入、修改、删除、更新、查询、统计、打印等，没有任何智能处理和灵气。如果计算机的数据处理停留于此，那它无非是起到了一个“大账本”加上一个“大算盘”的功能。人类对计算机的新追求是希望计算机能够具有人类的智能，从而可以代替人脑干更多的事情，成为名副其实的电脑。这就是以知识处理为标志的计算机应用的新阶段。

现代计算机理论的鼻祖冯·诺伊曼(John von Neumann)在病逝前完成了他的最后一部著作《计算机和人脑》(冯·诺伊曼 2010)。在这本小册子中，冯·诺伊曼在人们对电脑的一片赞誉声中冷静地指出：无论是从逻辑原理还是从算法结构上看，当时的电子计算机都与人脑迥然不同，记忆、存储方式也不一样，计算机远不及人脑。从那时至今，虽然从硬件上看计算机的发展经历了四代(电子管、晶体管、集成电路、

① 计算机指能进行数学运算的机器。有用机械装置做成的，如手摇计算机；有用电子元器件组成的，如电子计算机。现特指电子计算机，或称电脑，下同。

大规模集成电路),但从原理上看它们都还属于冯·诺伊曼式计算机。我们在期待新一代非冯·诺伊曼式计算机早日诞生的同时,也要努力让目前的计算机具有更多的知识处理能力。

与数值计算、数据处理相比,知识处理具有更大的复杂性和困难性。近四十多年来,以人工智能为代表的研究工作不断深入,并取得了一系列令人瞩目的成果(超级电脑"深蓝"打败国际象棋特级大师卡斯帕罗夫,各种机器人代替人类从事多种工作等)。多种非经典逻辑(多值逻辑、模糊逻辑、格值逻辑、模态逻辑、概率逻辑、非单调逻辑、自动认识逻辑等)的提出,多种软计算方法(模糊数学、神经网络、遗传算法、免疫算法、粗糙集方法、数据挖掘算法等)的诞生,为知识处理提供了新手段。

智能是人类所独有的能力,如何让计算机具有智能是一个极具挑战性的课题。谈到智能,离不开知识。而人类知识在计算机中如何表示?计算机如何获取这些知识?如何运用这些知识?这些重要的基本问题至今都没有得到权威的解答,因而仍在探索之中。

如果说计算机应用的前两个台阶(数值计算和数据处理)均已成功跨越,则知识处理这第三个台阶到目前为止也不过是刚刚起步。虽然数值计算和数据处理的成功为知识处理提供了基础和条件,但知识处理的广度、深度、难度和复杂度等都远非数值计算和数据处理所能比拟的。也许这些问题需要等到全新一代的非冯·诺伊曼式计算机问世后才能得到彻底解决。

本书的讨论局限于知识处理方面的一个十分确定、具体的问题:计算机学习问题,而且只介绍计算机学习领域中近年兴起的一种新方法——支持向量机(Support Vector Machines,简称 SVM)方法。本书特别聚焦于 SVM 的分类、回归方法,及其在天气预报业务中的应用。

SVM 方法思路巧妙,理论基础坚实,特别适用于处理高度非线性分类、回归等实际问题。与已有的其他方法相比,SVM 方法具有明显的优势,近年来在很多领域得到了成功的应用,在天气预报领域的初步应用也显示了很好的前景。

1.2 计算机学习的基本问题

学习是人类最重要的一种思维活动。无知的孩童通过不断地观察、学习和实践,日积月累地形成自己的思想,走向成熟,甚至发展成为杰出人才。

用人类的思维和实践揭开人脑学习的秘密是一个历史性的难题。如果说在几十年前它还只是少数哲学大师和学术先觉探讨的话题,那么由于信息技术的迅猛发展和计算机的广泛应用,学习问题,特别是计算机学习问题,已成为广大研究人员、技术人员必须面对的实际课题。

从逻辑的角度看,学习的本质是归纳,而学习所得结果的应用才是演绎。有人说,只凭推理演绎是得不出创造性的东西的,只有归纳才有可能创新。经典逻辑提供

了多种推理演绎的手段，但对于归纳学习却很少提出好办法。

人类的学习是一个复杂的思维过程，目前提出的计算机学习方法只是对这一过程的粗略模拟和近似。类似于从认识个体中抽象出概念，机器学习是通过已有样本数据集建立数学模型。

图 1.1 和 1.2 分别给出了这两种学习过程的粗略图示。

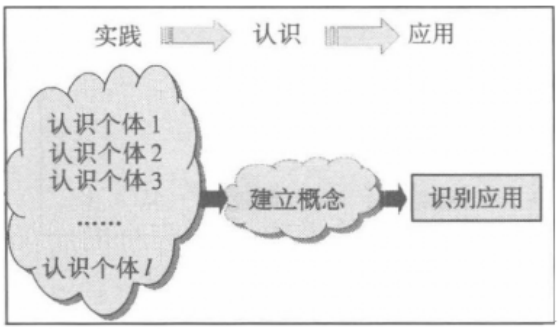


图 1.1　人类学习过程图示

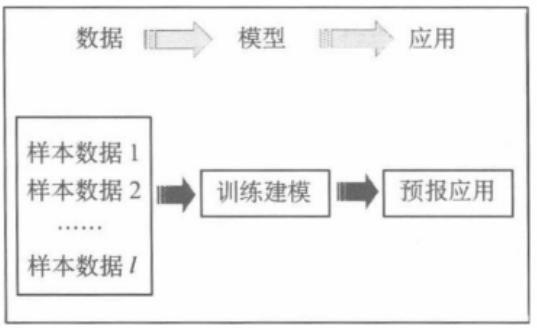


图 1.2　机器学习过程图示

比较两者的异同可以看出：

(1) 人脑从观察、实践中学习，感触到的大量事实是学习的素材；计算机从数据中学习，给定的相关数据集是学习的依据，通常称这样的数据集为训练集。

(2) 人脑学习的过程是通过对大量事实的分析、比较、对比、归纳等高级思维活动(甚至是形象思维)从而获得有用的知识；计算机学习是采用人类指定的某种确定算法对训练集进行“训练”学习，建立用于分类、回归等的数学模型。

(3) 人脑获得的知识可以用来指导以后的实践；机器学习获得的数学模型可用来对未来同类相关数据做识别、判断、推理、预测和预报。

(4) 人脑通过实践可以修正、提升已有的知识；计算机可以扩充训练数据集重新训练学习，改进已得的模型。

(5) 人脑的学习过程至今尚无方法被准确地形式化，它涉及复杂的生理、物理、化学、心理等过程；机器学习问题则是被人们完全形式化了的一个数学问题，它是对人脑学习过程的粗化和约简。

通俗地说，学习问题就是从有限数量的已经发生的历史样本数据中，寻找隐藏于其中的某种依赖关系，从而建立起可应用于未来预测预报的模型。或者用一句话概括：机器学习就是基于数据用计算机建模。

机器学习本质上就是一种统计方法。很多计算机对实际的应用问题的处理本质上都可以归结为上述机器学习问题。如通常所说的模式识别、回归分析、密度函数估计、聚类分析、人工神经网络等，都可以看成是这里所叙述的机器学习的特例。

图 1.3 和 1.4 分别给出了机器学习的直观描述和数据描述。

历史样本数据集由若干个成员组成，通常称一个成员为一个样本。因此观测数

据集也称样本集。每一个样本数据通常包含因子量(用 n 维向量 $\boldsymbol{X}$ 表示)和实况量(用实数 y 表示),因此历史样本数据集的第 i 个样本数据可以表示成($\boldsymbol{X}_i$, y_i)如图 1.4。其中 $\boldsymbol{X}_i=(x_{i1}, x_{i2}, \cdots, x_{in})$为 n 维向量,它对应于选取的预报因子的值;y_i 为预报量的实况值。

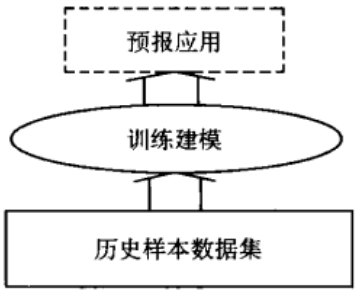


图 1.3 机器学习的直观描述

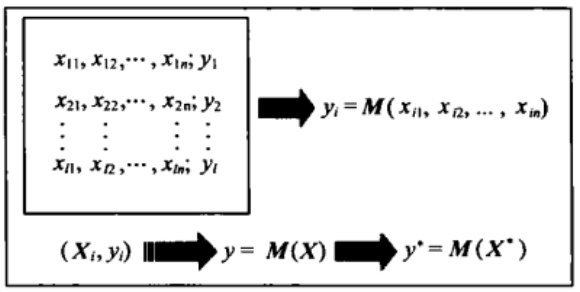


图 1.4 机器学习的数据描述

图 1.4 中所表示的 $y=\boldsymbol{M}(\boldsymbol{X})$是由样本数据集经过训练学习建立的模型。它是一个预报量与预报因子值的确定的函数关系,可用于预测预报。这一函数关系不仅应该同已有的历史样本数据充分符合(拟合),更重要的是要在对未来的预测预报进行应用时有良好的效果。

如果这一函数关系是线性函数,则对应的就是线性(分类或回归)模型,否则为非线性模型。本书称这样的一个函数关系为建立的一个学习机。

历史样本数据集是机器学习的基础,并不是随便凑成的一堆数据就可以称为样本数据集。必须指出,选取的预报因子和样本数据集的质量决定了建立的学习机是否有效。我们所说的样本数据集具有一定的客观性,它们是现实世界某一现象或过程的真实反映,但它们又具有一定的不确定性,往往是随机选取的。可以想象成是一只"上帝之手"把样本集的输入数据一一选取出来(我们相信数据集的生成是有规律的,但这个规律对我们来说是未知的)。而后它们通过一个完全确定过程的"黑箱"输出对应的数据,但该"黑箱"对于我们来说是未知的。机器学习的核心就是近似找出这一"黑箱"对应的输出输入关系。

用数学的语言来描述,图 1.3 可以细化为图 1.5:

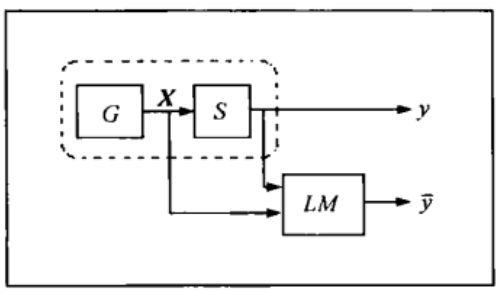


图 1.5 机器学习的图示

其中的框图 G 为输入随机向量(因子向量)产生器,所有输入数据看成是从满足某一固定但未知的概率分布函数 $F(\boldsymbol{X})$ 中独立生成的;框图 S 为产生器,它是一个黑箱系统,对每个输入向量产生一个确定的输出,产生输出的依据同样是客观存在且固定的,但对我们来说是未知的。一般认为这种依据是一个未知的条件概率分布函数 $F(y|\boldsymbol{X})$。产生器(G)和训练器(S)这两部分用虚线框起,它对应于实际问题的样本数据集。

我们称框图 LM 部分为一个学习机器,简称学习机。它通过对大量样本数据的训练学习,从指定的函数范围内确定一个数学模型(本书称建立一个学习机),作为对黑箱系统(训练器)的最好近似。所谓最好近似就是当把 G 产生的数据 $\boldsymbol{X}$ 同时输入到 S 和 LM 时,它们的输出值 y 和 $\bar{y}$ 最为接近。通常要预先指定待确定模型所属的函数类 $f(\boldsymbol{X},\boldsymbol{\alpha})$,从中选取一个最优函数 $f(\boldsymbol{X},\boldsymbol{\alpha}^*)$。这样,建立学习机的问题就变成确定最优函数所对应的一组参数 $\boldsymbol{\alpha}^*$ 的问题。

机器学习的过程就是由样本数据集建立学习机的过程。机器学习问题可以形式化为:

给定函数集 $f(\boldsymbol{X},\boldsymbol{\alpha})$,$\boldsymbol{\alpha}\in\boldsymbol{\Lambda}$ 和 l 个独立同分布的样本数据训练集

$$(\boldsymbol{X}_1,y_1),(\boldsymbol{X}_2,y_2),\cdots,(\boldsymbol{X}_l,y_l)$$

其中 $\boldsymbol{X}_i\in R^n$,为 n 维向量,如何从给定的函数集 $f(\boldsymbol{X},\boldsymbol{\alpha})$ 中选择出能够最接近训练器 S 响应的函数 $f(\boldsymbol{X},\boldsymbol{\alpha}^*)$来呢?

这里产生了如下一些理论问题:

(1) 学习机的备选函数类 $f(\boldsymbol{X},\boldsymbol{\alpha})$如何确定?范围太大(比如不加限制)无法计算,范围太小(比如限定线性函数)难以保证精度。

(2) 依据什么选取原则确定最优函数 $f(\boldsymbol{X},\boldsymbol{\alpha}^*)$?通常采用误差最小化原则,是否还有其他原则?相应的误差如何估算?

(3) 样本数据集是随机产生的,一般来说,即使是对于同一个实际预报问题,对应于不同样本的数据集所建立的学习机模型是不会完全一样的。但这些学习机随样本数据的增加是否有一个稳定的趋势,即收敛性?若收敛,收敛速度如何?可否调控?

(4) 建立的学习机即使对训练数据的拟合率很好,但推广能力也未必好。通过历史样本数据建立的学习机的泛化能力(即推广能力)如何估计?学习机的性能好坏如何评价?

Vapnik V N 教授创立的统计学习理论针对上述问题提出了一套理论成果。SVM 方法是统计学习理论的一种实用化实现,对上述问题,特别是问题(2)—(4),从算法到程序实现,都给出了有效可行的解决方案。讨论这些困难的理论问题不是本书的重点,只在第 6 章中结合统计学习理论做了简要介绍。

早期的机器学习研究几乎完全局限在人工智能这一狭窄领域中；现在，机器学习已经开始走进不同学科的不同领域，成为一种具有通用性的支撑和服务技术。而目前，SVM 方法仍然在所有的机器学习方法中具有明显的优势。

很多实际应用问题本质上都可以归结为上述机器学习问题。如通常所说的聚类分析、模式识别、回归分析、密度函数估计、人工神经网络等，都可以看成是这里所说的机器学习的特例。比如线性回归分析，就是在线性函数类中采用最小二乘法选取与样本点偏差平方和最小的线性函数。

1.3 SVM 方法的基本思想

模式识别、回归分析、密度函数估计是机器学习的三个基本内容。学者们应用概率统计和函数逼近等方法取得了很多研究成果，特别是在处理线性问题上，有些甚至可以说是完美的。但对于本质上是非线性的问题还缺少较好的结果。

当我们面对数据处理而又缺乏理论模型时，统计分析方法是最先被采用的方法。然而传统的统计方法只有在样本数量趋于无穷大时才能有理论上的保证。而在实际应用中样本数目通常都是有限的，甚至是小样本的，基于大数定律的传统统计方法对此难以取得理想的效果。

Vapnik 等人提出的统计学习理论是一种专门的小样本理论，它避免了人工神经网等方法出现网络结构难于确定、过学习、欠学习，以及局部极小等问题，被认为是目前针对解决小样本的分类、回归等问题的最佳理论。这一方法数学推导严密，理论基础坚实。基于这一理论近年提出的 SVM 方法，为解决非线性问题提供了一条新思路。

早在 20 世纪 60 年代，Vapnik 就已奠定了统计学习的基本理论基础，如经验风险最小化原则下统计学习一致性的条件（收敛性、收敛的可控性、收敛与概率测度定义的无关性，号称机器学习理论的“三个里程碑”）、关于统计学习方法推广性的界的理论，以及在此基础上建立的小样本归纳推理原则等。

直到 20 世纪 90 年代中后期，能够实现统计学习理论和原则的实用化算法——SVM 方法才逐渐被完整地提出，并且在模式识别等人工智能领域得到成功应用，受到了广泛关注。

SVM 方法的基本思想是：定义最优线性超平面，并把寻找最优线性超平面的算法归结为求解一个最优化（凸规划）问题。进而基于 Mercer① 核展开定理（Mercer 1909），通过非线性映射 φ，把样本空间映射到一个高维乃至于无穷维的特征空间

① Mercer，著名数学家，这里指的是他在 1909 年提出的一个重要定理，详见第 3 章 3.5 节。

(Hilbert① 空间),使在特征空间中可以应用线性学习机的方法解决样本空间中的高度非线性分类和回归等问题。简单地说就是实现升维和线性化。

降维(即把样本空间向低维空间做投影)是人们处理复杂问题常用的简化方法之一,这样做可以降低计算的复杂性。而升维映射,即把样本向高维空间做映射,大大增加样本的维数。通常这样做必然会增加计算的复杂性,甚至会引起"维数灾",因而人们很少问津。但是对于分类、回归等问题来说,很可能在低维样本空间无法进行线性处理的样本集,在高维特征空间却可以通过一个线性超平面实现线性划分(或回归)。

图 1.6 给出一个在二维空间中无法线性划分但映射到三维空间却可以线性划分的例子。映射是这样的:二维空间的样本点(x_1, x_2)映射成三维空间的点(x_1^2,$\sqrt{2}x_1x_2$, x_2^2)。两类样本点的在二维空间是线性不可分的,但在三维空间中却可以用一个平面把它们完全划分开。很多在低维空间中看似无法实现的"特异"事情,在高一维的空间中却是平常事。

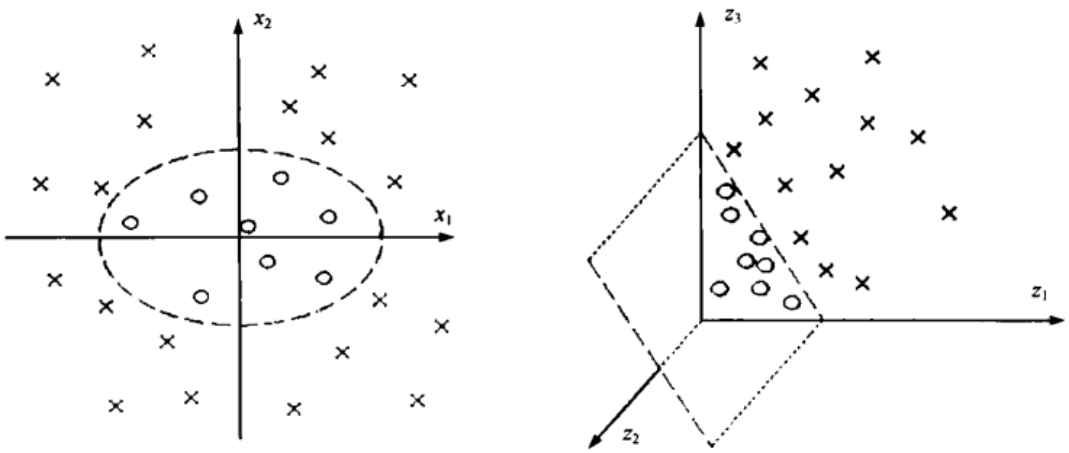


图 1.6　升维后实现线性可分的例子

SVM 的升维是通过一个非线性映射 φ,把样本空间中的点变换成一个称为特征空间的高维(甚至是无穷维)空间中的点,在特征空间中求解变换后的问题,这样做可以使原来难以解决的问题获得一个极为广阔的求解空间。在上面的例中,升维才只是增加一维:$\varphi(x_1,x_2)=(x_1^2,\sqrt{2}x_1x_2, x_2^2)$,在 SVM 方法中,维数通常增加很多,有时增加到无穷。

线性化方法是人们解决复杂问题的一种常用手段。线性问题算法简单,结果确定,容易解决。很多线性问题都有很完满的解。所谓线性化方法,就是把非线性问题

① Hilbert 空间是应用最广泛的一类泛函空间,它是完备的内积空间,有关内积空间的定义详见第 3 章 3.1.3 节。

近似、转化为线性问题解决，或借助于已解决的线性问题从而解决非线性问题。

SVM 方法的线性化是先求得(在高维空间中)线性分类或回归的解决办法，然后把该方法应用在变换后的高维特征空间中。在高维特征空间中得到的是问题的线性解，但与之相对应的却是原来样本空间中问题所寻找的非线性解。这样就借助解线性问题的方法解决了原来的非线性问题。图 1.7 对此给出了一种直观图示。

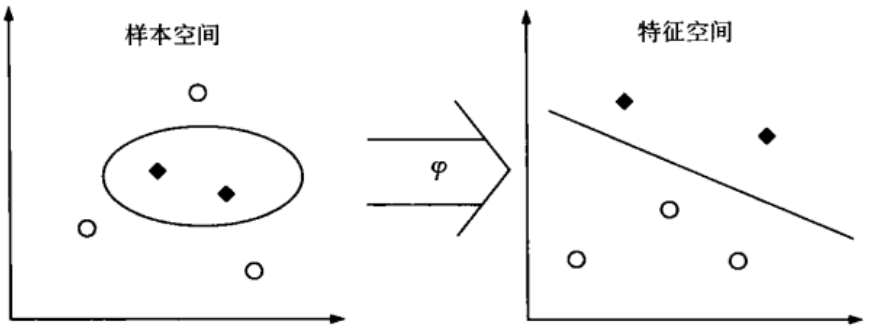


图 1.7 样本空间到特征空间的非线性映射示意图

一般的升维都会带来计算的复杂化。这里自然产生两个问题：一是如何找出非线性映射 φ 的数学表达式；二是如何简化升维带来的计算复杂性。

SVM 方法巧妙地解决了这两个难题：由于应用了核函数的展开定理，所以完全不需要知道非线性映射的显式表达式；由于是在高维特征空间中应用线性学习机的方法，所以与线性模型相比不但几乎不增加计算的复杂性，而且在某种程度上避免了“维数灾”。这一切要归功于核的展开和计算理论。因此人们又称 SVM 方法为基于核的一种方法。对于核方法的研究是比 SVM 更为广泛和深刻的研究领域。

1.4 SVM 方法的特点和应用展望

SVM 是一种有坚实理论基础的新颖的小样本学习方法。它基本上不涉及概率测度的定义及大数定律等，因此不同于现有的统计方法。从本质上看，它避开了从归纳到演绎的传统过程，实现了高效的从训练样本到预报样本的“转导推理”(transductive inference)，大大简化了通常的分类和回归等问题。与常规的统计方法相比，SVM 方法具有如下特点：

(1) SVM 的最终决策函数只由少数的支持向量所确定，计算的复杂性取决于支持向量的数目，而不是样本空间的维数，这在某种意义上避免了“维数灾”。如果说神经网络方法是对样本的所有因子加权的话，SVM 方法则对只占样本集少数的关键样本(支持向量)“加权”。当预报因子与预报对象间蕴涵的复杂非线性关系尚不清楚时，基于关键样本的方法可能优于基于因子的“加权”。

(2) 少数支持向量决定了最终结果，这不但可以帮助我们抓住关键样本、“剔除”

大量冗余样本，而且注定了该方法不但算法简单，而且具有较好的“鲁棒①”性。这里的鲁棒性主要是指在建模时增减非支持向量样本，其对建模结果没有影响。

(3) 由于有较为严格的统计学习理论作为保证，可以对用 SVM 方法建立模型的推广(泛化)能力作出评估。SVM 方法可以给出所建模型的推广能力确定的界估计(误差的上界，精度的下界)，这是目前其他任何学习方法所不具备的优势。

(4) 建立任何一个数据模型，人为的干预越少越客观。与其他方法相比，建立 SVM 模型所需要的先验干预较少。

(5) SVM 通过核函数实现低维到高维空间的非线性映射，所以适合于解决本质上非线性的分类、回归等问题。

近年来 SVM 方法已经在很多领域应用并取得了成功，显示了它的优势。气象预报预测领域的很多问题都带有显著的非线性特性，这一方法有望在释用数值预报产品、多项气象预报预测的研究和业务中得到推广应用。支持向量方法也为样本分析、因子筛选、信息压缩、知识挖掘和数据修复等提供了新工具。近年来 SVM 方法在分类预报和回归预报中的初步应用已经证明了这一点。

虽然 SVM 方法与基于概率测度和大数定律的传统统计方法风格迥异，但它与通常的统计分析预测方法有天然的联系(有大致相同的问题表述形式和数据预处理方法等)。对于应用了统计分析预测方法的其他研究和业务课题(诸如判别分析、主分量分析、时间序列分析等)，不妨用 SVM 方法一试并加以比较，相信会有新的发现。

Meyer D 等(2002)作为一个专门研究课题，将 SVM 方法与已有的其他机器学习方法作了系统严格地比较对照。

对于分类问题，他们将 SVM 方法与 16 种已有的分类方法作了比较。这 16 种方法涵盖了几乎目前所有的已知流行的分类方法，包括分类树方法、线性判别分析方法、平方判别分析方法、神经网络方法、最小近邻法等。效果比较中数据的选用是针对不同领域、不同类型的 21 个数据集进行的，其中包括医学数据、金融数据、生物数据、图形数据等。比较得出的结果是：对于分类问题，无论从分类精度、泛化能力、建模计算量等方面看，SVM 方法比其他方法都具有绝对的压倒优势。

对于回归问题，他们将 SVM 方法与 9 种已有的回归方法作了比较。这 9 种方法也涵盖了目前所有常用的回归方法。效果比较中数据的选用是针对不同领域、不同类型的 12 个数据集进行的。其中同样包括医学数据、金融数据、生物数据、图形数据等。比较得出的结果是：对于回归问题，SVM 方法具有明显的优势，对于特定数据类的优势突出。

① 鲁棒是 rubostness 的音译，也可意译为稳健性、健壮性等。

1.5 SVM 方法的参数优化

任何方法和技术都有弱点，即存在未解决的“开问题”，SVM 也不例外。SVM 方法又被称为“核方法”或“核机器”，这是因为核的展开和计算是这一方法的关键。这一方法目前存在的最大问题是核函数的选定及有关核参数的优化，这是目前尚未解决的问题，也是制约 SVM 方法应用的主要因素。至今人们还不能回答：对于特定的数据集，是否存在最优核函数和最优核参数？如果有，如何实现？

对于气象业务应用来说，我们希望并期待这一问题早日得到解决，但不可能等到所有理论问题都解决了之后才开始应用该方法。结合近年在天气预报中的应用经验，我们对这一问题作了多方面的研究，除取得了一些优选参数的经验方法外，还开发出一个实用化的参数寻优软件平台——全中文支持向量建模应用软件平台（Chinese Modeling for Support Vector Machines，CMSVM），其目前已开发出 2 个版本 CMSVM 1.0 和 CMSVM 2.0，它是应用 SVM 方法的得力工具，应用它可以大大减轻核参数寻优的负担。

在 CMSVM 2.0 中我们提出了一套实用化的“网格搜索”和“交叉验证”的参数寻优方法，虽然应用这一方法寻得的未必是理论上的最优，但它是满足一定条件下的“最优”。多年来在台站的推广应用的实践证明，该软件平台的功能较为齐备，运行稳定可靠，操作使用方便。该软件平台不但能适应目前气象业务应用的需要，而且已经引起很多其他行业（如计算机、遥感、水文、地质、地震、海洋、烟草、电力、农业、医学、经济和金融）、高等学校和科研院所科技人员和研究生的关注。

自 2004 年 SVM 的分类和回归方法被首次引入气象预报预测领域以来，该方法受到广大气象业务和科研工作者的关注，很快在晴雨预报、降水量预报、暴雨预报、面雨量预报、雷雨冰雹预报、台风路径预报、温度预报、天空云量预报、能见度和雾的预报、大气污染预报、水文气象预报、汛期旱涝预测、沙尘暴预测、场预测、短期气候预测，以及预报因子分析等广泛的领域得到实际应用。本书的附录列出了截至 2009 年底在相关杂志发表的应用论文（不完全统计）索引，它们刊载于《气象学报》、《应用气象学报》、《热带气象学报》、《气象》、《气象科技》、《干旱气象》、《暴雨灾害》等学术期刊。

1.6 本章小结

本章是对全书内容的概述。首先对信息技术的应用发展作了简略的历史回顾，从中引出计算机学习的基本问题和本书聚焦解决的具体问题。进而直观、通俗地介绍了 SVM 方法的基本思想和优缺点，并对其应用前景作了展望。

第 2 章　线性支持向量机模式识别

有人说科学始于分类。

“分类”严格的数学定义是指对给定集合元素的一个划分，它把原来的集合分成若干子集，使任何一个元素属于且只属于其中的一个子集。也就是说划分对于元素具有“不重不漏”的要求。经典数学中通常是先建立集合上的一个等价关系（满足自返性、对称性、传递性），再依据这一等价关系把集合分成一些等价子集（等价类）。这一方法多应用于理论研究中，对于实际应用中的具体问题往往难以找到分类的等价关系。

大量实际应用问题可以归结为分类问题，如天气预报、医疗诊断、地质勘探、管理决策等。这里涉及两个问题：一是用来分类的方法（又称模式）是否存在？如果存在，如何把它求出？二是把所建立的模式应用于对未来的元素进行分类是否有效？随着计算机和计算技术的发展，一个具有信息时代特色的新术语——模式识别应运而生。但从数学上看，模式识别就是分类的代名词。

人们已经提出了很多模式识别的方法。除非单纯为了进行比较对照，本书不涉及其他方法而只介绍 SVM 方法。

2.1　模式识别问题的表述

分类问题又称模式识别，就是依据有限的观测数据（训练样本）来寻求样本中蕴涵着的分类关系（建立分类模型），进而用求得的分类模型对未来数据（预报数据）进行预报。下面通过一个例子来说明这一过程。

例 2.1　诊断就诊人员是否患某种未知疾病 S 时通常要化验人体内各种元素（如锌、铁等）的含量。表 2.1 给出 90 个人的化验结果和诊断数据（数据未经过归一化处理），这组数据的前 30 个是已经确诊为患 S 的病人的化验结果；中间 30 个是已经确定为非患 S 的病人的化验结果；最后 30 个是待确诊的人的化验结果。

表 2.1　就诊人员检验数据和诊断结果

病人序号	是否确诊	锌	铜	铁	钙	锰	钾	钠
1	是	58.2	5.42	29.7	323	138	179	513
2	是	106	1.87	40.5	542	177	184	427
⋮	是	⋮	⋮	⋮	⋮	⋮	⋮	⋮
29	是	182	6.49	61.7	3870	432	143	367.5
30	是	235	15.6	23.4	1806	166	68.9	188
31	不是	166	15.5	24.5	700	112	179	513

续表

病人序号	是否确诊	锌	铜	铁	钙	锰	钾	钠
32	不是	185	15.7	31.5	701	125	184	427
⋮	不是	⋮	⋮	⋮	⋮	⋮	⋮	⋮
59	不是	90.0	3.27	8.17	622	52.3	770	852
60	不是	178	28.8	32.4	992	112	70.2	169
61	未知	213	19.1	36.2	2220	249	40.0	168
62	未知	170	13.9	29.8	1285	226	47.9	330
⋮	未知	⋮	⋮	⋮	⋮	⋮	⋮	⋮
89	未知	164	16.1	38.0	21.35	152	64.3	240
90	未知	179	21.0	35.0	1560	226	47.9	330

能否根据前 60 个人的这些数据特征和诊断结果，找出一种分类方法，初步判断后 30 名就诊人员是否患 S 病？

前 60 个数据是用来训练学习的数据，称为训练样本集，后 30 个数据为待预报数据集。

像例 2.1 这类模式识别问题可做如下数学描述：

给定一组训练样本集

$$(\boldsymbol{X}_1, y_1), (\boldsymbol{X}_2, y_2), \cdots, (\boldsymbol{X}_l, y_l) \tag{2.1}$$

其中 $\boldsymbol{X}_i \in R^n$，为 n 维向量，$y_i \in \{-1, 1\}$ 或 $y_i \in \{1, 2, \cdots, k\}$；$l$ 为训练样本的个数，k 为分类类数。

另外给出待预报数据集

$$\boldsymbol{X}_{l+1}, \boldsymbol{X}_{l+2}, \cdots, \boldsymbol{X}_{l+m} \tag{2.2}$$

寻求所对应的 y_i 输出值，m 为待预报样本个数。

模式识别的主要任务是通过对训练样本集数据的训练学习，建立一个分类模式 $y = M(\boldsymbol{X})$，使其不但能够对训练样本正确分类，而且具有较强的推广能力。

必须强调，单纯能够对训练样本正确分类是意义不大的，重要的是推广能力。所谓推广能力是指对于输入新的待预报数据 $\boldsymbol{X}_i$，可以由所建立的模式得到正确的对应输出值 y_i。有多种数学方法对有限数据可以完全拟合。在第 5 章的 5.4 节将给出一个可以对训练样本充分拟合但毫无推广预报能力的例子。当然，推广能力的好坏首先要依赖于数据本身的性质，这些数据是否反映了所研究过程的本质，是否具有代表性，其次才是所建立模型的好坏。为了与国际文献中常用的术语一致，以下称建立的分类模式 $y = M(\boldsymbol{X})$ 为学习机或分类机。

当 $y_i \in \{-1, 1\}$ 时，其为最简单的二类分类；当 $y_i \in \{1, 2, \cdots, k\}$ 时，其为 k 类分类。多类分类可看成是做多个二类分类（判断是否属于某一类）。因此以下我们将重

点讨论二类分类。

如果分类机 $y=M(\boldsymbol{X})$ 为线性函数(直线或线性超平面),则对应的模型为线性分类模型;否则为非线性分类。

线性分类的理想情况是训练样本集可以完全线性分离。当训练样本集本身不能线性分离(类有重叠现象)时,可以通过引入松弛变量转换为可线性分离的情况。图 2.1 给出了 SVM 的分类问题及其解决思路。

本章首先讨论线性可分离的二类分类的 SVM 分类机,进而解决线性不可分离的二类分类和多类线性分类问题。第 3 章将通过引入非线性映射和核函数,将复杂的非线性分类问题转化为线性 SVM 分类问题加以解决。

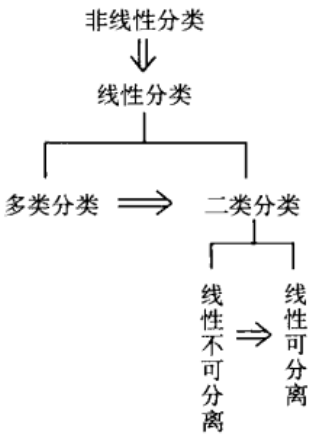


图 2.1　SVM 分类问题及解决思路

2.2　最优划分超平面与支持向量的概念

二类分类是比较常见的模式识别问题,即只有两个待识别的模式,分别称为正类和负类,如天气预报中的有雨和无雨预报。通常对于式(2.1)给定数据的二类分类,就是在规定的函数类中寻找一个合适的实函数 $y=f(\boldsymbol{X})$,使得对式(2.2)中的任何 $\boldsymbol{X}_i$,当 $f(\boldsymbol{X}_i)\geqslant 0$ 时把 $\boldsymbol{X}_i$ 归入正类,当 $f(\boldsymbol{X}_i)<0$ 时把 $\boldsymbol{X}_i$ 归入负类。总决策函数可以表示为

$$y = M(\boldsymbol{X}) = \operatorname{sgn}(f(\boldsymbol{X}))$$

在预报中 sgn 函数为符号函数,当它的自变量为非负数时取值+1,当自变量为负数时取值−1。

当 $f(\boldsymbol{X})$ 是线性函数时最好求解,称为二类线性分类问题。此时对于式(2.1)给出的训练样本集的线性二类分类问题,就是寻求分类学习机

$$\begin{aligned} M(\boldsymbol{X}) &= \operatorname{sgn}(w_1x_1 + w_2x_2 + \cdots + w_nx_n + b) \\ &= \operatorname{sgn}((\boldsymbol{W}\cdot\boldsymbol{X}) + b) \end{aligned} \tag{2.3}$$

使对于 $i=1,2,\cdots,l$ 满足条件

$$y_i = M(\boldsymbol{X}_i) = \operatorname{sgn}((\boldsymbol{W}\cdot\boldsymbol{X}_i) + b) \tag{2.4}$$

其中 $\boldsymbol{W}$, $\boldsymbol{X}_i$, $\boldsymbol{X}\in R^n$, $b\in R$,$\boldsymbol{W}$ 和 b 为待确定的参数,$\boldsymbol{X}_i$ 和 $\boldsymbol{X}$ 分别为训练样本和待分类样本数据,$(\boldsymbol{W}\cdot\boldsymbol{X}_i)$ 为通常的内积运算。与通常划分直线和划分平面的方法类似,我们称高维($n>3$)情况下的 $(\boldsymbol{W}\cdot\boldsymbol{X})+b=0$ 为划分超平面,$\boldsymbol{W}$ 为其法方向向量。满足条件式(2.4)意味着对于所有的样本点,应用得到的学习机做出的分类都正确。它又可以写成等价的简缩形式

$$y_i((\boldsymbol{W}\cdot\boldsymbol{X}_i)+b)\geqslant 0 \tag{2.5}$$

对于线性可分离的问题，形式与式(2.3)相同且满足式(2.5)条件的线性决策函数一般不是唯一的。图 2.2 给出二维情况下满足条件的划分直线的分布区域图。落在阴影区域内的任意一条直线都可作为决策函数。于是人们要问：两个不同的划分直线哪个预测效果更好？众多划分直线中是否有最优选择？若有，是否唯一？最优划分直线如何寻求？这一系列问题都可以推广到高维空间，只需要把划分直线替换成划分超平面就能实现。

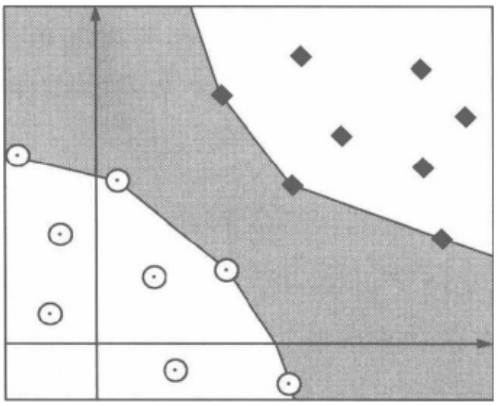


图 2.2 划分直线的分布区域

判断划分直线的优劣要有依据。通常采用误差最小化原则，即寻求决策函数使对训练样本集的分类误差“总和”(有多种汇总方法)最小。按此原则，落在阴影区域内的任意一条直线都是最优，因为它们都能使总分类误差为零。

直观上看，最优划分直线应该是对两类样本均是“公平”的、“不偏不倚”的，它应该在图 2.2 阴影区域的“中间”，使其“离”两类样本的“间隔”最大。这一直观想法可进行如下的数学描述：

定义 2.1 称 $\|\boldsymbol{X}_i-\boldsymbol{X}_j\|=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2}$ 为点 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 之间的距离，它是通常欧几里得空间中距离概念的推广。

定义 2.2 称 $\min\{\|\boldsymbol{X}-\boldsymbol{X}_i\|:\boldsymbol{X}\in R^n,(\boldsymbol{W}\cdot\boldsymbol{X})+b=0\}$ 为样本点 $\boldsymbol{X}_i$ 到超平面的距离，它是样本点 $\boldsymbol{X}_i$ 到超平面 $(\boldsymbol{W}\cdot\boldsymbol{X})+b=0$ 上所有点的距离最小者。

定义 2.3 称 $\min\{\|\boldsymbol{X}-\boldsymbol{X}_i\|:\boldsymbol{X}\in R^n,(\boldsymbol{W}\cdot\boldsymbol{X})+b=0,\ i=1,\cdots,l\}$ 为样本集到超平面的距离，它是样本集 $\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_l\}$ 所有样本点到超平面 $(\boldsymbol{W}\cdot\boldsymbol{X})+b=0$ 上所有点的距离最小者。

Vapnik(2004)提出一个间隔最大化(maximal-margin)原则。所谓间隔，是指训练样本集到划分超平面的距离。因而间隔最大化原则是指在对所有样本分类正确的条件下，寻求使间隔达到最大的划分。这就要对 $\boldsymbol{W}$ 和 b 寻优，以求得最大间隔

$$\max_{W,b}\{\min\{\| X - X_i \| : X \in R^n, (W \cdot X) + b = 0, i = 1, \cdots, l\}\}$$

定义 2.4　对应最大间隔的划分超平面称为最优划分超平面，简称为最优超平面。

图 2.3 给出二维样本的图示，其中两条平行虚线 l_1 和 l_2（称为边界）间距离之半就是最大间隔。图中的 L 为最优划分超平面。可以证明最大间隔是唯一的，但达到最大间隔的最优超平面可能不唯一。

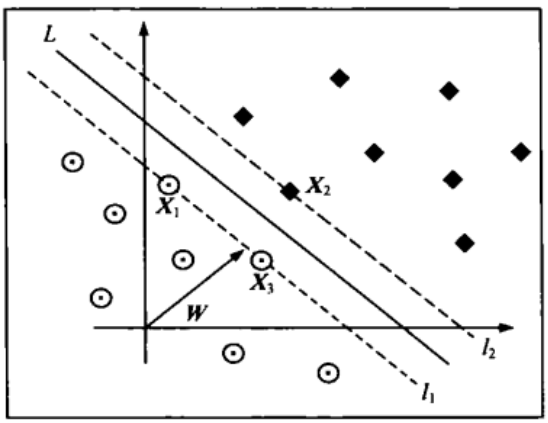


图 2.3　最优划分超平面示意图

最大间隔和最优超平面能够只由落在边界上的样本点完全确定，称这样的样本点为支持向量，如图 2.3 中的样本点 $\mathbf{X}_1$，$\boldsymbol{X}_2$ 和 $\boldsymbol{X}_3$。因此有：

定义 2.5　确定最优划分超平面的样本点称为支持向量。

在样本集线性可分离的情况下，支持向量都落在边界上。

支持向量是 SVM 方法中最核心的概念之一。如果能找出所有的支持向量，则可以完全确定最优超平面，从而完全确定分类模型。

由此，只由样本集中的少数训练样本点（支持向量）就把最大间隔和最优划分超平面完全确定，其余非支持向量的样本点均不起作用，这一点具有十分重要的意义。它说明在间隔最大化原则下的最优划分不是依赖于所有点，而只由支持向量所决定。求最优划分超平面和最大间隔等同于确定各个样本点是否为支持向量。这预示着该方法具有好的鲁棒性和简洁的算法。

由支持向量确定的线性分类机称为线性 SVM。

2.3　最优划分超平面的求解

对于式(2.1)给定的训练样本集，如果训练样本是线性可分离的，如何建立线性 SVM？依据 Vapnik 提出的间隔最大化原则，可以把求解最优超平面的问题归结为一个最优化问题。

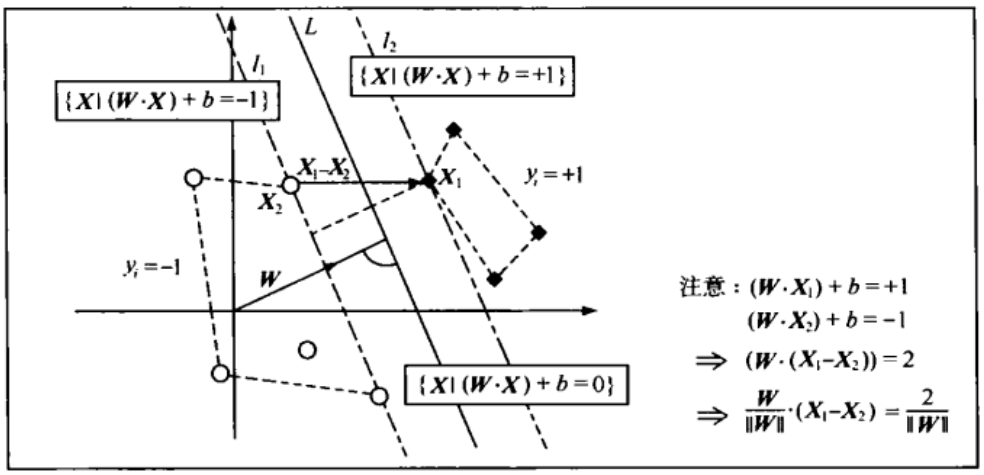


图 2.4 最优划分超平面示意图

设图 2.4 中的划分超平面 L 的方程为

$$(\boldsymbol{W}\cdot\boldsymbol{X})+b=0$$

则两条边界 l_1，l_2 的方程(经过恒等变形后)为

$$(\boldsymbol{W}\cdot\boldsymbol{X})+b=\pm 1$$

设 $\boldsymbol{X}_1$ 在 l_2 上，$\boldsymbol{X}_2$ 在 l_1 上，即

$$(\boldsymbol{W}\cdot\boldsymbol{X}_1)+b=+1$$

$$(\boldsymbol{W}\cdot\boldsymbol{X}_2)+b=-1$$

两式相减有

$$[\boldsymbol{W}\cdot(\boldsymbol{X}_1-\boldsymbol{X}_2)]=2$$

进而有

$$\left[\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|}\cdot(\boldsymbol{X}_1-\boldsymbol{X}_2)\right]=\frac{2}{\|\boldsymbol{W}\|} \tag{2.6}$$

式(2.6)左边恰好就是连接 $\boldsymbol{X}_2$，$\boldsymbol{X}_1$ 两点的向量在划分超平面法方向上的投影①，它是最大间隔的二倍。因此，求最大间隔等价于求 $\|\boldsymbol{W}\|$ 的最小值。为计算方便，它等同于求 $\|\boldsymbol{W}\|^2$ 或 $\frac{1}{2}\|\boldsymbol{W}\|^2$ 的最小值。

① 根据两个向量点积的定义 $\boldsymbol{A}\cdot\boldsymbol{B}=\|\boldsymbol{A}\|\,\|\boldsymbol{B}\|\cos\alpha$，若向量 $\boldsymbol{A}$ 为单位向量，即 $\|\boldsymbol{A}\|=1$，则 $\boldsymbol{A}\cdot\boldsymbol{B}=\|\boldsymbol{B}\|\cos\alpha$，$\alpha$ 是 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 的夹角，所以 $\|\boldsymbol{B}\|\cos\alpha$ 为向量 $\boldsymbol{B}$ 在 $\boldsymbol{A}$ 上的投影。$\boldsymbol{A}$ 相当于这里的 $\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|}$，$\boldsymbol{B}$ 相当于 $(\boldsymbol{X}_1-\boldsymbol{X}_2)$。

考虑到要使所有训练样本点分类正确且满足间隔要求，所以最小化要满足条件

$$若\ y_i = 1，则(\boldsymbol{W} \cdot \boldsymbol{X}_i) + b \geqslant 1$$

$$若\ y_i = -1，则(\boldsymbol{W} \cdot \boldsymbol{X}_i) + b \leqslant -1$$

两式可以合并为

$$y_i((\boldsymbol{W} \cdot \boldsymbol{X}_i) + b) \geqslant 1$$

这样，就将建立线性 SVM 的问题转化为求解如下一个规划问题

$$\min\left\{\frac{1}{2}\|\boldsymbol{W}\|^2\right\}$$

$$约束条件：y_i((\boldsymbol{W} \cdot \boldsymbol{X}_i) + b) \geqslant 1 \qquad (i = 1, \cdots, l) \tag{2.7}$$

注意：这里的目标函数是凸函数，约束条件构成一个凸集。这样的规划问题称为凸规划。

对于凸规划，在最优化理论中有一个著名的 Kuhn-Tucker 定理：

定理 2.1　设 $f(\boldsymbol{X})$为连续的凸函数，$g_i(\boldsymbol{X})$为线性函数。对于如下定义在 R^n 上的最优化问题

$$\min\{f(\boldsymbol{X})\} \qquad (\boldsymbol{X} \in R^n)$$

$$约束条件：g_i(\boldsymbol{X}) \geqslant 0 \qquad (i = 1, \cdots, l)$$

其存在一个最优点 $\boldsymbol{X}^*$ 使 $f(\boldsymbol{X})$达到最小值的充分必要条件是：存在非负的 $\boldsymbol{\alpha}^*$ ($\boldsymbol{\alpha}^* \geqslant \boldsymbol{0}$)，使得下式成立

$$\begin{cases} \dfrac{\partial L(\boldsymbol{X}^*, \boldsymbol{\alpha}^*)}{\partial \boldsymbol{X}} = 0 \\ \dfrac{\partial L(\boldsymbol{X}^*, \boldsymbol{\alpha}^*)}{\partial \boldsymbol{\alpha}} = 0 \\ g_i(\boldsymbol{X}^*) \geqslant 0 \qquad (i = 1, \cdots, l) \\ \alpha_i^* g_i(\boldsymbol{X}^*) = 0 \qquad (i = 1, \cdots, l) \end{cases}$$

这里的 $L(\boldsymbol{X}, \boldsymbol{\alpha}) = f(\boldsymbol{X}) - \sum_{i=1}^{l} \alpha_i\ g_i(\boldsymbol{X})$ 为拉格朗日 (Lagrange) 函数，$\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)$，$\alpha_i^*$ 为 Lagrange 系数。最后一式称为 KKT 条件 (Karush-Kuhn-Tucker Condition)，它意味着 α_i^* 和 $g_i(\boldsymbol{X}^*)$ 至少要有一个为零，这表明不在约束边界上的点对应的 α_i^* 必都为零 (因为 $g_i(\boldsymbol{X}^*) \neq 0$)。

我们不加证明地直接引用上述定理求解最优化问题式 (2.7)。由于目标函数和

约束条件都满足 Kuhn-Tucker 定理，因而最优化问题式(2.7)存在唯一全局最小解。

因为求解过程需要运用最优化和对偶规划等较多数学方法，为方便应用，这里只给出求解的思路，而略去了具体的推导过程。

根据定理 2.1，最优化问题式(2.7)还应补充一个 KKT 约束条件

$$\alpha_i(y_i((\boldsymbol{W}\cdot\boldsymbol{X}_i)+b)-1)=0 \qquad (i=1,2,\cdots,l) \tag{2.8}$$

求解问题式(2.7)首先需要定义一个拉格朗日函数，如下：

$$L(\boldsymbol{W},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{W}\|^2-\sum_{i=1}^{l}\alpha_i(y_i((\boldsymbol{W}\cdot\boldsymbol{X}_i)+b)-1) \tag{2.9}$$

对 L 关于 b 求偏导数并令偏导数为零，则得到

$$\sum_{i=1}^{l}\alpha_i y_i=0 \tag{2.10}$$

再对 L 关于 $\boldsymbol{W}$ 求偏导数并令偏导数为零，经整理后得到

$$\boldsymbol{W}=\sum_{i=1}^{l}\alpha_i y_i\boldsymbol{X}_i \tag{2.11}$$

将式(2.10)和(2.11)代入式(2.9)经过化简整理，就消去了 $\boldsymbol{W}$ 和 b，使 $L(\boldsymbol{W},b,\boldsymbol{\alpha})$ 变成了 $J(\boldsymbol{\alpha})$，

$$\begin{aligned}J(\boldsymbol{\alpha})&=\frac{1}{2}\Big(\sum_{i=1}^{l}\alpha_i y_i\boldsymbol{X}_i\cdot\sum_{j=1}^{l}\alpha_j y_j\boldsymbol{X}_j\Big)+\sum_{i=1}^{l}\alpha_i-\Big(\sum_{i=1}^{l}\alpha_i y_i\boldsymbol{X}_i\cdot\sum_{j=1}^{l}\alpha_j y_j\boldsymbol{X}_j\Big)-b\sum_{i=1}^{l}\alpha_i y_i\\&=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\Big(\sum_{i=1}^{l}\alpha_i y_i\boldsymbol{X}_i\cdot\sum_{j=1}^{l}\alpha_j y_j\boldsymbol{X}_j\Big)\\&=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{\substack{i=1\\j=1}}^{l}\alpha_i\alpha_j y_i y_j(\boldsymbol{X}_i\cdot\boldsymbol{X}_j)\end{aligned}$$

这样就把原来的问题式(2.7)转化为它的一个对偶规划问题如下

$$\begin{aligned}&\max J(\boldsymbol{\alpha})=\max\Big\{\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{\substack{i=1\\j=1}}^{l}\alpha_i\alpha_j y_i y_j(\boldsymbol{X}_i\cdot\boldsymbol{X}_j)\Big\}\\&\text{约束条件：}\sum_{i=1}^{l}\alpha_i y_i=0 \qquad (\alpha_i\geqslant 0;i=1,\cdots,l)\end{aligned} \tag{2.12}$$

比较最优化问题式(2.7)和它的对偶问题式(2.12)，可以看出：

(1)原问题式(2.7)是将问题最小化，对偶问题式(2.12)是将问题最大化。这可以粗略解释为：在拉格朗日函数 $L(\boldsymbol{W},b,\boldsymbol{\alpha})$，即式 (2.9)中，最优解对应于一个“马鞍

点”，它是关于 $\boldsymbol{W}$ 的最小，同时又是关于 $\boldsymbol{\alpha}$ 的最大。有关对偶规划的具体理论请参看相关书籍。

(2)原问题式(2.7)所要求出的 $\boldsymbol{W}$ 的分量个数为 n，与样本点的维数一样。而对偶问题式(2.12)所要求出的系数 $\boldsymbol{\alpha}$ 的分量个数为 l，与样本点的个数一样多，而与样本向量的维数无关。这在一定意义上避免了数值计算中讨厌的“维数灾”，也使我们可以无所顾忌地采用“升维”的手法。

对于解决问题式(2.12)，已有很多成熟的数值计算方法，可以解出对应的最优解 $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)$ 来。其实，其中非支持向量对应的 α_i^* 都为零。

将 $\boldsymbol{\alpha}^*$ 代入式(2.11)，可以求得 $\boldsymbol{W}^*$；将已知的 $\boldsymbol{W}^*$ 代入 KKT 条件式(2.8)可以求得

$$b^* = -\frac{1}{2}[\boldsymbol{W}^* \cdot (\boldsymbol{X}_+ + \boldsymbol{X}_-)],$$

其中 $\boldsymbol{X}_+$ 和 $\boldsymbol{X}_-$ 分别为正类和负类中的任意一个支持向量。

将它们代回最优超平面方程 $(\boldsymbol{W} \cdot \boldsymbol{X}) + b = 0$ 中得到

$$M(\boldsymbol{X}) = \operatorname{sgn}((\boldsymbol{W}^* \cdot \boldsymbol{X}) + b^*) = \operatorname{sgn}\Big(\sum_{\text{支持向量}} \alpha_i^* y_i (\boldsymbol{X}_i \cdot \boldsymbol{X}) + b^*\Big)$$

从而得到最优超平面决策函数为

$$M(\boldsymbol{X}) = \operatorname{sgn}\Big(\sum_{\text{支持向量}} \alpha_i^* y_i (\boldsymbol{X} \cdot \boldsymbol{X}_i) + b^*\Big) \tag{2.13}$$

其中 α_i^* 和 b^* 为确定最优划分超平面的参数，$(\boldsymbol{X} \cdot \boldsymbol{X}_i)$ 为两个向量的内积。

这样，求解最优划分超平面的问题就通过求解一个对偶最优化问题而得以解决。根据凸规划理论，此方法得到的解是全局最优解。

由式(2.8)可知：非支持向量对应的 α_i 都为零，因而式(2.13)中的求和只对少数支持向量进行，即最优决策函数只取决于支持向量。这也给出了支持向量的另一个定义：

定义 2.6　不为零的 α_i 对应的样本点称为支持向量。

从求解过程可以发现：最优化问题式(2.7)中只有少数约束条件(由支持向量对应)对最终结果起作用，其余约束条件可以拿掉而不影响结果。这种看似奇怪的现象却并不难于理解，举一个形象的例子：把一个小球放入一个封闭立方体的盒子里求最优位置(势能最小)，盒子对小球有六个约束面，小球最后可能停靠在某一角(被三个面约束)，而没有约束它的三个面是否存在都不影响它的最优位置。

Smola J 等(1998b)还对这一全局最优性进一步给出了一种力学解释：

可以将最优超平面想象成一块刚性板(如图 2.5 中的 L)：

每一个支持向量 $\boldsymbol{X}_i$ 都对 L 施加一个作用力 $\boldsymbol{F}_i$，其方向与超平面的法向量为同向

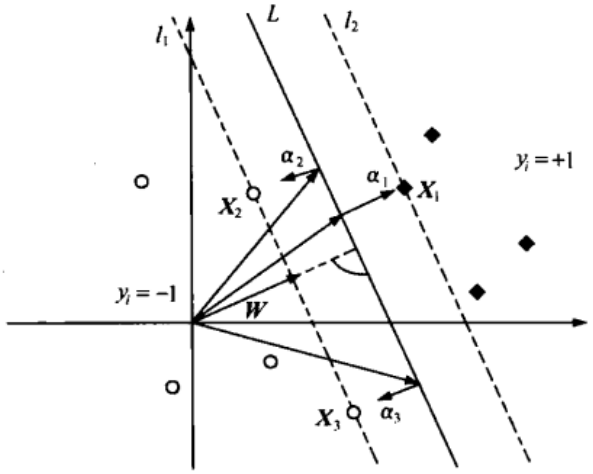


图 2.5 最优划分超平面的物理解释

平行力，方向分别指向它所在的类，力的大小为 α_i，则可以表示成

$$\boldsymbol{F}_i = \alpha_i y_i \boldsymbol{W}_0 \qquad (\boldsymbol{W}_0 = \boldsymbol{W}/\|\boldsymbol{W}\|),$$

结合式(2.10)说明作用在 L 上的力满足同向平行力的平衡条件，即合力为零。

如果把 $\boldsymbol{X}_i$ 看成是以坐标原点为起点，对应的样本点为终点的矢量，则式(2.11)说明所有支持向量的作用力矢量关于坐标原点的合力矩为零，这是因为：

$$\sum_{i=1}^{l} \boldsymbol{X}_i \times \boldsymbol{F}_i = \sum_{i=1}^{l} \boldsymbol{X}_i \times (\alpha_i y_i \boldsymbol{W}_0) = \sum_{i=1}^{l} (\alpha_i y_i \boldsymbol{X}_i) \times \boldsymbol{W}_0 = \boldsymbol{W} \times \boldsymbol{W}_0 = 0。$$

这样，我们从物理的角度又对支持向量有了新的理解：所有样本点中只有支持向量才对最优超平面提供力与转矩，且合力与合转矩恰好达到力学稳态平衡。

在线性 SVM 模式识别方法中，只通过少数训练样本点就确定了最优划分超平面，其余样本均不起作用。划分不是依赖于所有点，而只由支持向量决定。这说明这一方法不但具有很好的“鲁棒性”，而且在数据处理、个例分析、因子分析等方面有重要意义。

2.4 线性不可分问题的求解

所谓“线性不可分离”的情况，是指实际上根本不存在线性超平面可以把两类样本完全正确划分开。当我们面对一个新的实际问题时，必须先从假定它是线性不可分离开始计算。对于式(2.1)那样给定的训练样本集，如果训练样本是线性不可分离的，即样本集中的两类样本间有交叠，其根本不存在能完全正确分类的划分超平面，这表现为约束条件：$y_i((\boldsymbol{W} \cdot \boldsymbol{X}_i)+b) \geqslant 1$ $(i=1,\cdots,l)$ 可能永远得不到满足，当样本数量较大时，这种情况经常发生，而绝大多数实际问题都是这样。造成这种情况的原因可能有：

(1) 高噪音：个别样本数据可能是“干扰”样本，难以剔除；

(2) 问题本身就不是完全线性的，只能近似看成线性。

此时可否建立 SVM 线性划分模型？

注意，这里的前提是我们还要“勉为其难”地建立线性 SVM。因此这只能允许有错分的情况出现，但希望错分的样本尽量少，被错分的样本的偏差尽量小，要在允许错分的情况下寻求最优。比如，如果说分错就要“罚款”的话，希望总“罚款”最少。类似于引入惩罚机制，通常通过引入线性规划中的松弛变量(slack variables)来实现。

所谓松弛变量，就是对每一个样本点 $\boldsymbol{X}_i$ 引入一个非负的变量 $\xi_i(\geqslant 0)$，它是被错分的样本点到对应边界超平面的垂直距离(ξ_k)，相当于错分样本的“罚款”数。当 $\boldsymbol{X}_i$ 被正确分类(在两条边界线之外)时，则 ξ_i 等于零；且 $\boldsymbol{X}_i$ 被错分的偏差越大，ξ_i 值就越大。

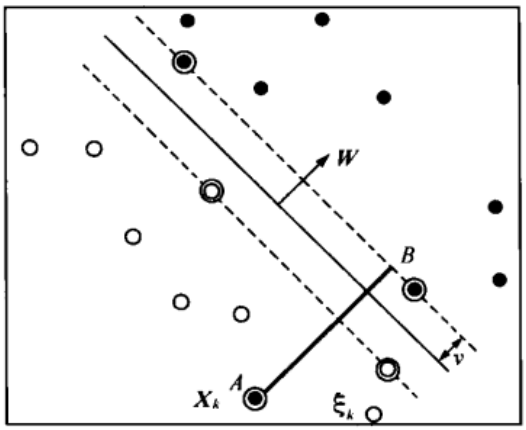


图 2.6　松弛变量几何示意图

图 2.6 给出了松弛变量(ξ_k)的几何解释。其中样本 $\boldsymbol{X}_k$ 被错误分类，它到自己应在的类的边界的距离(即线段 AB 的长度)，就是它的松弛变量值 ξ_k。

ξ_i 的值与样本点的位置、分类超平面位置和间隔 ν 有关，在叠代寻优的数值计算时，它们是这样计算的

$$\xi_i = \xi[(\boldsymbol{X}_i, y_i), (\boldsymbol{W}, b), \nu] = \max\{0, \nu - y_i((\boldsymbol{W} \cdot \boldsymbol{X}_i) + b)\}$$

在叠代寻优计算时，对于固定的超平面($\boldsymbol{W}$, b)、固定的间隔 ν，ξ_i 可取不同值：当 $\xi_i = 0$ 时，意味着对 $\boldsymbol{X}_i$ 的分类正确；当 $0 < \xi_i \leqslant \nu$ 时，意味着 $\boldsymbol{X}_i$ 落在了正确的分类方向，但在边界线以外；当 $\xi_i > \nu$ 时，说明对 $\boldsymbol{X}_i$ 进行的分类出现了错误，它到自己应在的类边界的距离为 ξ_i。

这样，对于式(2.1)给出的样本集求最优划分超平面，要兼顾两个方面：不但要考虑使划分边界最大，还要同时考虑使总偏差(罚款)最小。这可以通过修改式(2.7)中的目标函数和约束条件，应用完全类似的方法进行求解。与式(2.7)类似的新的凸规

划问题为

$$\min\left\{\frac{1}{2}\|\boldsymbol{W}\|^2+C\sum_{i=1}^{l}\xi_i\right\}$$

$$\text{约束条件：}y_i((\boldsymbol{W}\cdot\boldsymbol{X}_i)+b)\geqslant 1-\xi_i \qquad (i=1,\cdots,l) \tag{2.14}$$

$$\xi_i\geqslant 0 \qquad (i=1,\cdots,l)$$

若 ξ_i 都为零，式(2.14)就变成了线性可分问题式(2.7)；而式(2.14)中大于零的 ξ_i，其对应的样本被错分(这里的错分指超越了应在类的边界线)；参数 C 为惩罚系数，其越大说明对于错分的惩罚越严厉。

从数值计算上看，求解问题式(2.14)与求解式(2.7)本质上是一样的，同样是凸约束下的凸规划。利用定理 2.1，同样消去 $\boldsymbol{W}$，b 和 $\boldsymbol{\xi}$，式(2.14)则转化为对偶问题式(2.15)

$$\max J(\alpha)=\max\left\{\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{\substack{i=1\\j=1}}^{l}\alpha_i\alpha_j y_i y_j(\boldsymbol{X}_i\cdot\boldsymbol{X}_j)\right\}$$

$$\text{约束条件：}\sum_{i=1}^{l}\alpha_i y_i=0 \qquad (i=1,\cdots,l) \tag{2.15}$$

$$0\leqslant\alpha_i\leqslant C$$

同样求解式(2.15)得到的最优超平面决策函数仍然为

$$M(\boldsymbol{X})=\operatorname{sgn}\Big(\sum_{\text{支持向量}}\alpha_i^* y_i(\boldsymbol{X}\cdot\boldsymbol{X}_i)+b^*\Big) \tag{2.16}$$

注意：在问题式(2.14)中出现的 C 为“惩罚系数”，它表示对于错分的惩罚力度，C 越大对错分的惩罚越严厉；在实际计算的对偶问题式(2.15)中，它又是支持向量系数 α_i 的上界。从物理上看这是合理的：对错分样本的惩罚力度不能小于正确分类样本对分类超平面的“贡献”。事实上，落在边界上的支持向量(当 $\xi_i=0$ 时)的系数 α_i 小于 C，被错分的支持向量(当 $\xi_i>0$ 时)对应的系数 α_i 等于 C，而其他样本的 α_i 为零。

2.5 线性多类分类问题的求解

在很多实际问题中，待识别的模式并不是两个而是多个，如降水预报中的分级预报等。这也对应于多类分类问题。如果分类函数是线性的，则称为线性多类分类问题。图 2.7 给出一种三类线性分类的示意图。

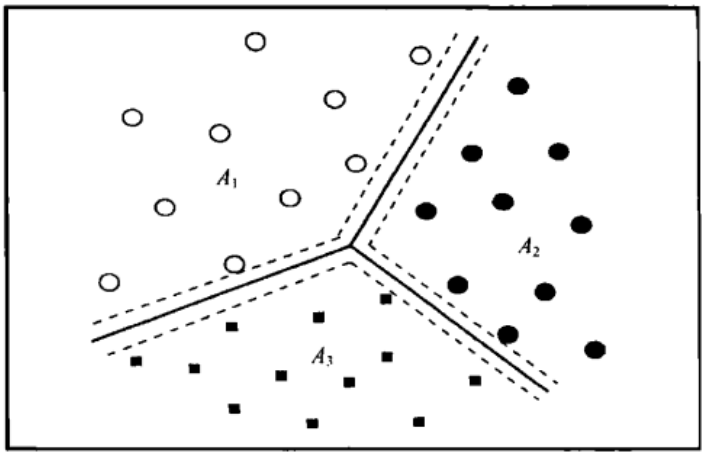


图 2.7　三类线性分类示意图

二类线性分类是多类线性分类的基础。很多解决多类线性分类问题的方法，都是将其转化为多个二类线性分类问题来解决。原则上讲，如果一种二类线性分类的方法应用效果不佳，那么基于这种方法的多类线性分类的效果只会更差。下面介绍三种将多类线性分类问题转化为多个二类分类问题的方法和直接进行多类分类的方法。

2.5.1　“一对其余”方法

将 K 类分类问题转化为 K 个二类分类问题，其中每个二类分类都是判断样本点是属于第 i 类或是其余类。

对应于每一个二类分类的决策函数为

$$f_i(\boldsymbol{X}) = \operatorname{sgn}((\boldsymbol{W}_i \cdot \boldsymbol{X}) + b_i) \qquad (i = 1,2,\cdots,K),$$

其含义为

若 $\operatorname{sgn}((\boldsymbol{W}_i \cdot \boldsymbol{X}) + b_i) = 1$，则 $\boldsymbol{X}$ 属于第 i 类；

若 $\operatorname{sgn}((\boldsymbol{W}_i \cdot \boldsymbol{X}) + b_i) = -1$，则 $\boldsymbol{X}$ 不属于第 i 类(即属于其余类)。

最后总的决策函数为

$$M(\boldsymbol{X}) = \arg\max((\boldsymbol{W}_i \cdot \boldsymbol{X}) + b_i) \quad 1 \leqslant i \leqslant K \tag{2.17}$$

上式的 arg 为选取指标函数，argmax()为选取使括号中的值达到最大的下标的值。这里是取使$(\boldsymbol{W}_i \cdot X + b_i)$达到最大的 i 的值，其含义为：选取使样本点 $\boldsymbol{X}$ 对于 K 个决策函数值中最大的 $f_i(\boldsymbol{X})$的指标 i 作为 $M(\boldsymbol{X})$的值。i 所对应的类，就是样本点 $\boldsymbol{X}$ 应该归属的类。

这样，K 类分类就可以通过进行 K 个二类分类，再加上一个选择决策从而得到

解决。选择决策是一种“投票机制”，它有不同的实现途径。式(2.17)给出的是最简单、最常用的一种解决方法。

例 2.2 图 2.7 给出的是三类划分问题，可以归结为建立三个二类分类的 SVM M_1、M_2 和 M_3(如图 2.8)。通过比较待预报样本在 M_1、M_2 和 M_3 三个 SVM 中的分类值大小，来判断待预报样本所在的类别。

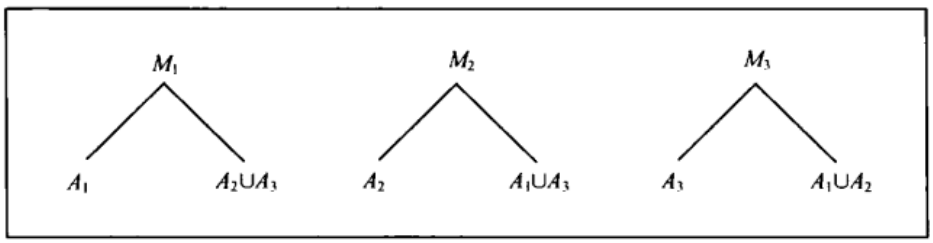


图 2.8 “一对其余”三类分类的集合示意图

这种方法可以充分利用训练样本数据建立模型(全部样本参加每一个二类分类的建模)，但每个模型的训练集都要单独标定。当不同类别间的特征差异明显时，运用该方法进行分类的效果较好。

2.5.2 “一对一”方法

这是另一种通过多个线性二类分类实现 $K(K>2)$类线性分类的方法。具体做法是：将训练样本按类分成 K 个不同的子集。

K 类中任取两类，如第 i 和第 j 两类，做一次二类线性分类，求得一个二类分类函数：

$$f_{ij}(\boldsymbol{X}) = \operatorname{sgn}((\boldsymbol{W}_{ij} \cdot \boldsymbol{X}) + b_{ij}) \qquad (i = 1, \cdots k;\ j = 1, \cdots, k;\ i > j),$$

其含义为

若 $\operatorname{sgn}((\boldsymbol{W}_{ij} \cdot \boldsymbol{X}) + b_{ij}) = 1$，则 $\boldsymbol{X}$ 属于第 i 类；

若 $\operatorname{sgn}((\boldsymbol{W}_{ij} \cdot \boldsymbol{X}) + b_{ij}) = -1$，则 $\boldsymbol{X}$ 属于第 j 类。

对于 K 类分类，需要建立 $K(K-1)/2$ 个二类分类函数。最后总的决策函数通过引入投票机制来决定，比如

$$M(\boldsymbol{X}) = \arg\max\{(\boldsymbol{W}_{ij} \cdot \boldsymbol{X}) + b_{ij}\} \qquad 1 \leqslant i, j \leqslant K,\ i > j \tag{2.18}$$

与“一对其余”相比，当 $K>3$ 时，“一对一”方法需要建立更多的学习机。此外，“一对一”方法对于训练样本数据集的要求较高，每次建模都是只有部分数据进入建模，当训练样本数目较少或类分布不平衡时，该方法的应用有较大局限性。

无论是“一对其余”还是“一对一”方法，两种方法都过度依赖于“投票机制”，都可能存在分类的“死区”问题。以三类线性分类为例(如图 2.9)，样本空间可能被三个

线性分类函数 L_1,L_2 和 L_3 切割成 7 个区域。

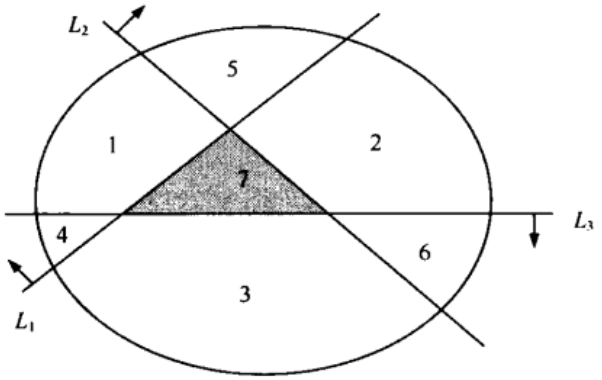


图 2.9　三类分类“死区”问题示意图

图 2.9 中 1,2 和 3 是理想区域,这些区域的点对于三个分类函数来说,只有一个取正值,其余为负值,这说明落在这些区域的样本点有完全确定的归属——归属于取正值的类。其余四个区域都没有这么“理想”,比如区域 7 里的点对应的所有决策函数均为负值,如果某个样本点 $\boldsymbol{X}$ 落入区域 7,这意味着它不属于三类中的任何一类,此时被迫依据“投票机制”[如式(2.17)所给出的决策函数]进行分类判决则有些勉强,有时效果并不好。区域 4,5 和 6 也有类似问题。

2.5.3　逐步二分类方法

如果多类线性分类的类之间具有某种序关联,采用逐步二分类方法可以避免“投票机制”。所谓类之间的序关联,比如像降水预报中的大雨、中雨和小雨,空气质量预报中的重度污染、中度污染和轻度污染等,不同的类均可依据某一物理量的大小来划分。

逐步二分类方法就是先把所有样本分成两个子类,做一次二类划分;再对每个子类重复这一过程,直到得到的子类都是单一的类为止。对于 K 类线性分类问题,采用逐步二分类方法需要做$(K-1)$个二类线性分类。

这种方法实际上是一种二叉决策树方法。它之所以能避免“投票机制”,是因为最后的决策是通过顺序调用从上至下的有关决策函数来实现。需要注意的是对于同一个问题,决策树可能不唯一。

2.5.4　直接多类 SVM 分类

近年来应用 SVM 方法做多类分类的理论研究有很多报道。较早提出的最直接的方法是把多类问题归结为如下的一个统一的凸规划问题进行求解

$$\min\varphi(\boldsymbol{W},\boldsymbol{\xi}) = \frac{1}{2}\sum_{m=1}^{K} \| \boldsymbol{W}_m \|^2 + C\sum_{i=1}^{l}\sum_{m\neq y_i}\xi_i^m$$

$$\text{约束条件：}\ y_i((\boldsymbol{W}_{y_i}\cdot\boldsymbol{X}_i)+b_{y_i})\geqslant(\boldsymbol{W}_m\cdot\boldsymbol{X}_i)+b_m)+2-\xi_i^m$$

$$\xi_i^m\geqslant 0(i=1,\cdots,l;m\in\{1,2,\cdots,K\}\backslash y_i)$$

其中 l 为样本个数，K 为类数。

通过求解一个最优化问题，同时求得 K 个决策函数，但最后决策函数中仍需引入投票机制，如

$$M(\boldsymbol{X})=\arg\max((\boldsymbol{W}_i\cdot\boldsymbol{X})+b_i)\qquad(1\leqslant i\leqslant K)。$$

这一方法看似简单，实则大大增加了算法的复杂性，至今无法实际应用。

2.5.5 四种方法的比较

前面提到的四种多类分类方法各有利弊，表 2.2 从需要建立的学习机个数、训练样本集的要求、算法的复杂性及是否需要投票机制四个方面对它们进行了比较。

表 2.2 比较四种多类分类方法

K 类方法	学习机个数	训练样本集	算法复杂性	是否需要投票机制	效果
一对其余	K	全体样本	较好	是	一般
一对一	$K(K-1)/2$	约 $2/K$ 样本	较好	是	一般
逐步二分类	$K-1$	逐渐变小	较好	否	较好
直接多类 SVM	1	全体样本	差	是	一般

2.6 本章小结

（1）依据间隔最大化原则，通过定义最优划分超平面和支持向量，把线性可分的两类分类问题转化成一个凸约束条件下的凸规划问题。利用已有的最优化理论和成熟算法，建立起线性 SVM，使线性可分离问题得以解决，而且得到的是全局最优解。

（2）通过引入松弛变量，修改线性可分离的凸规划问题的目标函数和约束条件，把两类样本重叠的线性不可分问题用同处理线性可分离问题类似的方法解决，建立起一般的线性 SVM 模型。

（3）本质上看，多类划分是多个二类划分的综合。本章简单介绍了四种线性多类分类方法，并对其加以分析比较，从而把二类分类的线性 SVM 方法推广应用到对线性多类划分问题的处理。

第3章　非线性支持向量机模式识别

线性分类问题是模式识别中最简单的情况，这类问题已经有较多办法可以对其进行处理，效果还是令人满意的。真正的线性分类问题在现实中是不存在的，就像现实中并不存在严格数学意义上的直线和平面一样。大量的实际问题都是本质上非线性的，或带有非线性特征，分类问题也不例外。如图 3.1 所示，分类边界不是一条直线（或超平面），而是一些难以用解析函数公式表示出的复杂曲线。对此类情况进行分类至今仍具有很大的挑战性，人们一直没有找到很好的解决办法。

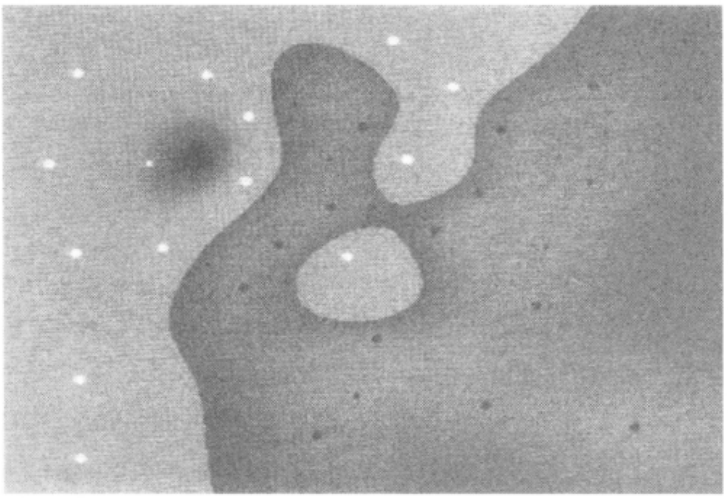


图 3.1　二维数据非线性划分图示

第 2 章介绍的 SVM 方法如果只能够处理线性分类问题，那它就不会受到人们的重视。事实上正是这样，前面提到的主要结果在四十多年前 Vapnik 就提出了，但并未引起人们的关注。只是近年来由于引入核方法，从而在应用上有所突破，可以用 SVM 来解决非线性问题，而且效果优于以往的大多数方法，才使 SVM 成为机器学习的主流方向之一。更重要的是 SVM 方法有坚实的理论基础，它可以对所建立模型的推广能力的界给出估计。

SVM 方法能够用来有效地解决非线性问题，这是其真正有价值的应用。如何从线性 SVM“升级”到非线性 SVM？简单地说就是通过一个非线性映射 φ，把样本空间映射到一个高维乃至于无穷维的特征空间，使原来的非线性问题在高维特征空间中变为线性可解的问题，从而可以在高维特征空间中应用线性 SVM 方法，解决样本空间中的高度非线性分类问题。幸运的是：求解过程并不需要知道非线性映射 φ 的显式表达式，似乎是“虚晃一招”，问题就迎刃而解了！

3.1 数学预备知识

本节我们暂时离开模式识别问题，而是以当前中国高校理工本科的高等数学基础为起点，把理解本章内容所必需的基础知识（如泛函等）先集中起来，作一简略通俗（但不具备数学的严格性）的介绍。

n 维向量 $\boldsymbol{F}$ 和在有限区间上定义的一元函数 $f(x)$ 是大家熟悉的。两者看似不同，其实，函数和向量本来就有着密切的关系：一个函数在数值计算时必须离散化，其结果就表现为一个向量；而一个向量当维数无限增加时就表现为一个实函数。难怪数学家们把函数也称为向量并建立了泛函分析理论。粗糙地说，当向量的维数增加到无穷时，向量就变成了一元函数；反之，一个一元函数在数值计算时进行离散化就变成了一个向量。图 3.2 对这一理论给出了相应的说明图示。

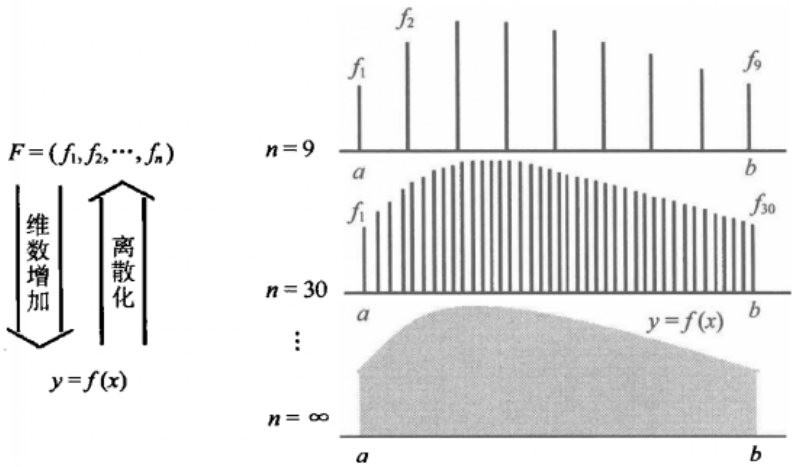


图 3.2 向量与函数的对应关系

同样，矩阵 $A_{n\times n}$ 与有限区域上定义的一个二元函数 $K(x,y)$ 也有类似的对应。当矩阵的维数增加到无穷时，它就变成一个二元函数；反之，一个二元函数离散化就变成了一个矩阵。图 3.3 对这一联系给出了相应的说明图示。

基于这种对应，人们就可以很容易地把线性代数中的一系列结果推广到函数空间中。

3.1.1 核与积分变换

可积函数定积分的计算结果为一个实数，但如果让被积函数乘上一个二元函数再做积分，则积分结果为一个新的函数

$$\int_a^b K(x,y)\varphi(x)\mathrm{d}x = \psi(y) \tag{3.1}$$

如果把 $\varphi(x)$ 看作输入，$\psi(y)$ 看作输出，式(3.1)给出了一种如图 3.4 所示的函数

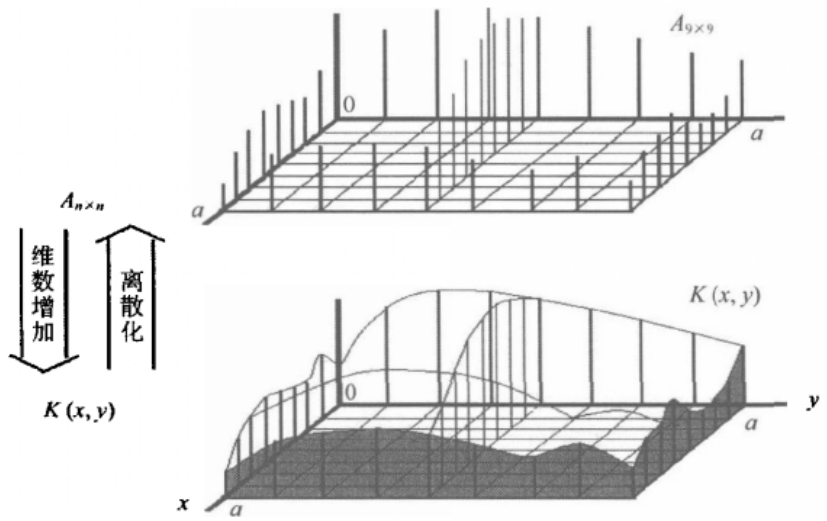


图 3.3　矩阵与二元函数的对应

变换，通常称为积分变换。

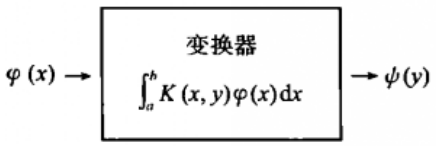


图 3.4　积分变换图示

定义 3.1　式(3.1)中的二元函数 $K(x,y)$通常被称作是一个积分变换的核函数，简称为核。

不同的核对应于不同的积分变换，如常见的傅立叶变换、拉普拉斯变换等。

定义 3.2　如果核 $K(x,y)$对于两个自变量 x 和 y 是对称的，即对任意 x 和 y 成立 $K(x,y)=K(y,x)$，则称核 $K(x,y)$为对称核。

定义 3.3　给定核 $K(x,y)$，若对于任意函数 $\varphi(\cdot)$都成立

$$\int_a^b\int_a^b K(x,y)\varphi(x)\varphi(y)\mathrm{d}x\mathrm{d}y>0,$$

则称 $K(x,y)$为正定核。与此类似还可定义非负定核($\geqslant0$)、负定核(<0)等。

核是研究积分方程和积分算子中引出的一个概念。核 $K(x,y)$可以是复函数。但需要注意的是，本书所有讨论只限定于实函数核，并且限定它对于两个自变量是连续的。注意：核中的自变量可以推广为向量 $\boldsymbol{X},\boldsymbol{Y}$。下面我们要用到的都是实对称正定核。

核在数值计算时需要被离散化，离散化后的结果是形成一个 n 行 n 列的方阵 $\boldsymbol{A}$。这告诉我们：核是实矩阵向无穷维的推广，同样对称核、正定核分别是实对称矩阵、正

定矩阵的自然延伸和推广，与之对应的矩阵定义分别是 $\boldsymbol{A}=\boldsymbol{A}'$ 和 $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}>0$，这里 $\boldsymbol{A}'$ 和 $\boldsymbol{X}'$ 分别代表矩阵 $\boldsymbol{A}$ 和向量 $\boldsymbol{X}$ 的转置。

进一步分析比较还会发现：式(3.1)给出的积分变换实际上是矩阵的线性变换 $\boldsymbol{A}\boldsymbol{X}'=\boldsymbol{Y}$ 的自然推广。只不过矩阵的线性变换是向量(数组)到向量的变换，而这里是函数到函数的变换。可以看出：实对称正定核是实对称正定矩阵的推广。

3.1.2 核的特征值与特征向量

矩阵除了有阶数、秩等指标外，最重要的指标就是方阵的特征值和特征向量：给定 n 阶方阵 $\boldsymbol{A}$，若有实数 λ 和非零向量 $\boldsymbol{X}'$，使得成立 $\boldsymbol{A}\boldsymbol{X}'=\lambda\boldsymbol{X}'$，则称实数 λ 为 n 阶方阵 $\boldsymbol{A}$ 的一个特征值，称非零向量 $\boldsymbol{X}'$ 为 n 阶方阵 $\boldsymbol{A}$ 对应于特征值 λ 的一个特征向量。

特征向量的特点是原向量经过线性变换后不改变方向，方向与原向量在同一直线上，而只是长度伸缩了 λ 倍。

与此类似还可以给出核的特征值与特征函数的定义：

定义 3.4 给定核 $K(x,y)$，若有 $\lambda\neq0$ 和 $\psi(x)$ 使得成立 $\int_a^b K(x,y)\psi(x)\mathrm{d}x=\lambda\psi(y)$，则称 λ 为核的一个特征值，称 $\psi(x)$ 为核 $K(x,y)$ 的关于特征值 λ 的一个特征函数。

与正定矩阵一样，正定核的特征值都是正实数。核的特征值与特征函数是矩阵的特征值和特征向量的自然推广。

有些核具有无穷多个特征值与特征函数，这些特征函数构成了一类重要的特殊函数集。

3.1.3 内积与内积空间

n 维向量空间理论中有一个重要结果：任何 n 维向量空间都可以找到一组(n 个)特殊的单位向量，两两彼此正交，称为一组标准正交基。该向量空间中的任何一个向量，都可以通过这组标准正交基线性表示出，且表示方法唯一。线性表示的系数称为向量的坐标，这可以将对向量的讨论转为对坐标的讨论。

如何将这些结果推广到一般空间？关键是需要在一般空间中定义内积。定义了内积，就可以把直观欧氏空间的结构和结论移植到高维函数空间。限于篇幅，这里只能进行粗略介绍。

定义 3.5 给定函数空间 $\mathscr{X}$ 和 $\mathscr{X}\times\mathscr{X}$ 到 $\boldsymbol{R}$ 上的一个映射 $\cdot$：

$$\cdot:\mathscr{X}\times\mathscr{X}\to\boldsymbol{R}$$

如果对于任意 $\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}\in\mathscr{X}$ 和 $\alpha\in\boldsymbol{R}$ 满足下列条件，就称之为一个内积：

(1) $(\boldsymbol{X}\cdot\boldsymbol{X})\geqslant0$，当且仅当 $\boldsymbol{X}=0$ 时 $(\boldsymbol{X}\cdot\boldsymbol{X})=0$ (非负性)

(2) $(\boldsymbol{X}\cdot\boldsymbol{Y})=(\boldsymbol{Y}\cdot\boldsymbol{X})$ (对称性)

(3) $((\boldsymbol{X}+\boldsymbol{Y})\cdot\boldsymbol{Z})=(\boldsymbol{X}\cdot\boldsymbol{Z})+(\boldsymbol{Y}\cdot\boldsymbol{Z})$；$(\alpha\boldsymbol{X}\cdot\boldsymbol{Y})=\alpha(\boldsymbol{X}\cdot\boldsymbol{Y})$ (拟线性)

定义 3.6　定义了内积的线性空间$\mathcal{X}$称为内积空间。

在内积空间中向量 $\boldsymbol{X}$ 的长度可以表示为：$\|\boldsymbol{X}\|=\sqrt{(\boldsymbol{X}\cdot\boldsymbol{X})}$，长度又称范数；两个向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的距离可以表示为 $\|\boldsymbol{X}-\boldsymbol{Y}\|$，两个向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间夹角余弦值可以表示为$(\boldsymbol{X}\cdot\boldsymbol{Y})/(\|\boldsymbol{X}\|\ \|\boldsymbol{Y}\|)$。因而，当$(\boldsymbol{X}\cdot\boldsymbol{Y})=0$ 时，称两个向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 正交，记作 $\boldsymbol{X}\perp\boldsymbol{Y}$。正交的概念可以推广到向量与集合、集合与集合。

定义 3.7　设$\mathcal{X}$为一个内积空间，若$\mathcal{X}$的一个子集 $\boldsymbol{S}$ 中的所有元素之间均是两两正交的，就称 $\boldsymbol{S}$ 为一个正交集；若 $\boldsymbol{S}$ 中每个元素的范数都是 1，那么对于所有 $\boldsymbol{X},\boldsymbol{Y}\in\boldsymbol{S}$，则有

$$(\boldsymbol{X}\cdot\boldsymbol{Y})=\begin{cases}0 & \text{当 } \boldsymbol{X}\neq\boldsymbol{Y}\\ 1 & \text{当 } \boldsymbol{X}=\boldsymbol{Y}\end{cases}$$

此时，称 $\boldsymbol{S}$ 为$\mathcal{X}$的一个标准正交集。

对于标准正交集，有如下的展开关系：

设 $\boldsymbol{S}=\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_k,\cdots\}$为$\mathcal{X}$的一个标准正交集，则对于任意 $\boldsymbol{X}\in\mathcal{X}$存在唯一的展开式

$$\boldsymbol{X}\sim\sum_i\alpha_i\boldsymbol{e}_i,$$

其中 $\alpha_i=(\boldsymbol{X}\cdot\boldsymbol{e}_i)$。

对于满足一定条件的$\mathcal{X}$和 $\boldsymbol{S}$，上面的展开式变成等式，这时我们称标准正交集 $\boldsymbol{S}$ 为空间$\mathcal{X}$的一组标准正交基。

例 3.1　设$\mathcal{X}$为$[0,2\pi]$上的可积函数空间，对于任意 $\boldsymbol{X},\boldsymbol{Y}\in\mathcal{X}$定义内积为

$$(\boldsymbol{X}\cdot\boldsymbol{Y})=\int_0^{2\pi}f(t)g(t)\mathrm{d}t,$$

这里 $\boldsymbol{X}=f(x)$，$\boldsymbol{Y}=g(x)$，因而，容易验证函数集：

$$\{1,\sin x,\sin 2x,\cdots,\sin nx,\cdots,\cos x,\cos 2x,\cdots,\cos nx,\cdots\}$$

就是一个正交函数集。乘以适当系数后，这个函数集就变成$\mathcal{X}$的一个标准正交集，而且是一组标准正交基，因为$\mathcal{X}$中的任意一个元素 $f(x)$(函数)，都可以由这个函数集唯一准确的线性表示出，表出式就是该函数展开的傅立叶级数。

例 3.2　设$\mathcal{X}$为$[a,b]$上的可积函数空间，任意 $\boldsymbol{X},\boldsymbol{Y}\in\mathcal{X}$类似上例定义内积(积分限改成 a 到 b)。设 $K(x,y)$ 为一实对称核，它的对应于特征值 λ_i 的特征向量为 $\psi_i(x)$。则函数集$\{\psi_i(x)\mid i=1,2,\cdots\}$为一个正交函数集。通过施密特正交化方法可以使它成为标准正交函数集。

常用的标准正交函数集还有勒让德(Legender)多项式、厄米特(Hermite)多项式

和拉盖尔(Laguerre)多项式等。

3.1.4 Mercer 核和 Mercer 定理

线性代数中实对称正定矩阵具有很好的性质,因而受到重视,同样实对称正定核也是十分重要的。

定义 3.8 对称正定的连续核称为 Mercer 核。

关于这类核有下面这一重要定理:

定理 3.1 Mercer 核函数 $K(x,y)$可以展开成一致收敛的函数项级数(Mercer J 1909)

$$K(x,y) = \sum_i \lambda_i \psi_i(x)\psi_i(y) \tag{3.2}$$

其中 λ_i,$\psi_i(x)$分别为核 $K(x,y)$的特征值和特征函数,它们的个数可能有限或无穷。

注意:Mercer 核的特征值都是正数。它的不同特征值对应的不同特征函数彼此则构成了一个正交函数集。式(3.2)又称为核的展开式或表现式,核可以通过自身的特征函数来进行表示,或者说 Mercer 核就是自身所有特征的"综合体",这充分体现了"特征"的字面含义。

上述定理当年在积分方程、积分算子研究中曾起过重要作用,人们很难想到,近百年以后它竟成了新兴的机器学习方法的重要理论基石。

3.2 非线性映射与特征空间

SVM 方法真正有价值的应用是用来解决非线性问题。当实际问题本质上是非线性时,线性 SVM 就无法建立,或者勉强建立起来了但应用效果也很差。如何用 SVM 方法解决非线性分类问题?

我们称讨论原问题所在的论域为样本空间,记为$\mathscr{X}$,通常它就是 $\boldsymbol{R}^n$。当在$\mathscr{X}$中不能进行线性分类时,可以考虑引入一个非线性映射 φ,把样本空间映射到一个新的空间$\mathscr{F}$,称为特征空间(图 3.5)。显然,特征空间$\mathscr{F}$**是由** $\mathscr{X}$中点的像 $\varphi(\mathscr{X})$所组成。由于$\mathscr{F}$是新空间,与原问题对应的新问题完全有可能在$\mathscr{F}$中通过线性方法得以解决。

这种"腾挪"的方法是人们的常用手段之一。比如,通过取对数可以把乘除运算变成加减运算,通过傅立叶变换可以使函数的微积分运算变成函数的乘除运算等。采用这种转换手法效果的好坏关键在于所选取的映射 φ。

在能够保持现有有用信息的前提下,以往人们往往关注于通过降维映射使问题简化。如例 3.3:

例 3.3 质点由空间位置(三个坐标)和质量所确定。当计算两点间的引力时,需要考虑八维样本向量 $\boldsymbol{X}=(x_1,y_1,z_1,x_2,y_2,z_2,m_1,m_2)$给解决问题带来的难度。为了使问题降维,引入映射:$R^8 \to R^3$

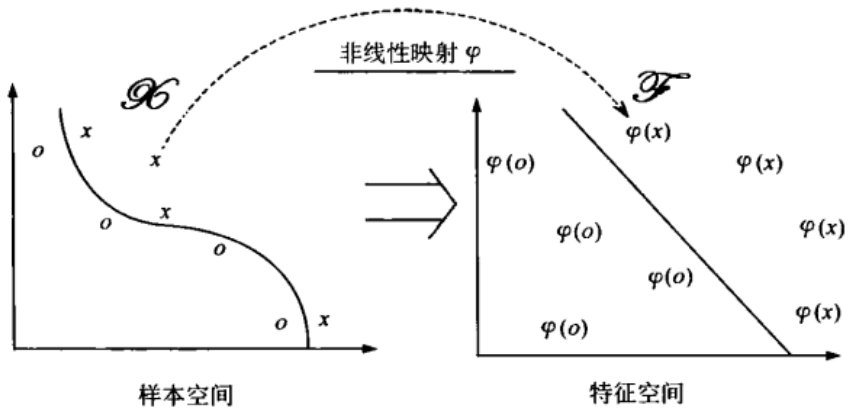


图 3.5　样本空间到特征空间的非线性映射

$$\boldsymbol{X}=(x_1,y_1,z_1,x_2,y_2,z_2,m_1,m_2)\mapsto\varphi(\boldsymbol{X})=(r,m_1,m_2),$$

其中，$r=\sqrt{\sum(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$。这样处理并没有丢失与计算引力有关的信息，却使问题得到了简化。

降维映射本质上是丢失了信息，通常这样的映射是不可逆的，因此它的适用范围很有限。因而，升维映入了人们的视线。

当人们把目光投向使向量维数增加的映射时，首先就是担心计算的复杂性：会不会带来所谓的“维数灾”？这种自然的担心甚至掩盖了增维会带来的很多令人惊异的结果。

一般来说，在低维空间中无法实现的事情，在高维空间中也许易如反掌地就能实现，这样的例子举不胜举。升维映射本质上是增加信息，或者说是使一些隐藏的信息凸显出来。

我们举一个升维的简单例子，它正好也对应于后面要提到的多项式核。

线性模型只用到样本数据的一次项，最简单的非线性函数就是高次多项式。第 1 章中的图 1.6 就是一例，那里只用到了 2 次项。如果样本点 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$ 的全部信息包含在 $x_1,x_2,\cdots,x_n$ 的直到 d 次方的项中，我们可以把其中的样本点映射到一个高维特征空间$\mathscr{F}$中去，这个特征空间包括 $x_1,x_2,\cdots,x_n$ 的从零次直到 d 次的项（也包含交叉项），经计算得知$\mathscr{F}$的维数是

$$C_{n+d}^{d}=\frac{(n+d)!}{d!n!}。$$

例 3.4　当 $n=3,d=2$ 时，上述特征空间的向量维数是 10，映射为

$$\varphi(x_1,x_2,x_3)=(x_1^2,x_2^2,x_3^2,\sqrt{2}x_1x_2,\sqrt{2}x_1x_3,\sqrt{2}x_2x_3,\sqrt{2}x_1,\sqrt{2}x_2,\sqrt{2}x_3,1)。$$

当 n 和 d 变大时，特征空间$\mathscr{F}$的维数呈“爆炸”式地增长。比如，对于十分粗糙的图像信息，每幅图像具有 $n=16\times16=256$ 个像素，样本空间的向量为 256 维。当 d

取 5 时，特征空间的向量维数则约为 97 亿！现有的计算机几乎无法直接对其计算和建模。

如果在算法上没有什么“绝招”，这种特征空间再广阔、再美好，也是无人敢问津的，更不用说是无穷维了。幸运的是由于核函数的引进使我们避免了这种维数灾。

构造特征空间的非线性升维映射绝不限于幂函数。比如可以用例 3.1 给出的正交函数集做映射处理

$$\boldsymbol{X} \mid\rightarrow \varphi(\boldsymbol{X}) = (1, \sin\boldsymbol{X}, \sin 2\boldsymbol{X}, \cdots, \sin n\boldsymbol{X}, \cdots, \cos\boldsymbol{X}, \cos 2\boldsymbol{X}, \cdots, \cos n\boldsymbol{X}, \cdots)。$$

如果例 3.2 中的核为 Mercer 核，则可以用其特征值和特征函数做如下映射

$$\boldsymbol{X} \mid\rightarrow \varphi(\boldsymbol{X}) = (\sqrt{\lambda_1}\psi_1(\boldsymbol{X}), \sqrt{\lambda_2}\psi_2(\boldsymbol{X}), \cdots, \sqrt{\lambda_k}\psi_k(\boldsymbol{X}), \cdots) \tag{3.3}$$

由于 Mercer 核的特征值都是正的，所以上面的根式是有意义的。式(3.3)就是我们下面要用到的重要映射关系式。

3.3 特征空间中的线性学习机

在第 2 章的内容里，我们已经在样本空间$\mathscr{X}$中建立起了线性学习机。假设通过一个非线性映射 φ 把$\mathscr{X}$映射到了特征空间$\mathscr{F}$，它是一个高维或无穷维空间(升维)。进而我们可以把第 2 章的同样方法应用到特征空间上，在特征空间上进行线性划分(线性化)。所以只要将样本空间的线性分类机中的 $\boldsymbol{X}$ 和 $\boldsymbol{X}_i$ 分别用 $\varphi(\boldsymbol{X})$和 $\varphi(\boldsymbol{X}_i)$替换，就可以很容易地建立起特征空间上的线性分类机。

对于特征空间上的线性可分离情况，问题归结为

$$\begin{aligned} &\min\left\{\frac{1}{2}\|\boldsymbol{W}\|^2\right\} \\ &\text{约束条件：} y_i((\boldsymbol{W}\cdot\varphi(\boldsymbol{X}_i))+b) \geqslant 1 \qquad (i = 1, \cdots, l) \end{aligned} \tag{3.4}$$

同样利用定理 2.1 的推导思路消去 $\boldsymbol{W}$ 和 b，由此式(3.4)化为对偶问题式(3.5)

$$\begin{aligned} &\max J(\boldsymbol{\alpha}) = \max\left\{\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{\substack{i=1\\j=1}}^{l}\alpha_i\alpha_j y_i y_j(\varphi(\boldsymbol{X}_i)\cdot\varphi(\boldsymbol{X}_j))\right\} \\ &\text{约束条件：} \sum_{i=1}^{l}\alpha_i y_i = 0 \qquad (\alpha_i \geqslant 0;\ i = 1, \cdots, l) \end{aligned} \tag{3.5}$$

最后求得最优超平面决策函数为

$$M(\boldsymbol{X}) = \operatorname{sgn}\Big(\sum_{\text{支持向量}}\alpha_i^* y_i(\varphi(\boldsymbol{X})\cdot\varphi(\boldsymbol{X}_i)) + b\Big) \tag{3.6}$$

对于特征空间上的线性不可分及多类问题都可以以类似方法进行处理，这里略

去。

必须明确：式(3.6)给出的是特征空间上的线性划分超平面，但与之相对应的却是实现了样本空间上的高度非线性划分，图 3.6 给出了相应的图示，其中的三条曲线实际上分别对应于特征空间中的最优划分直线和两条边界线。

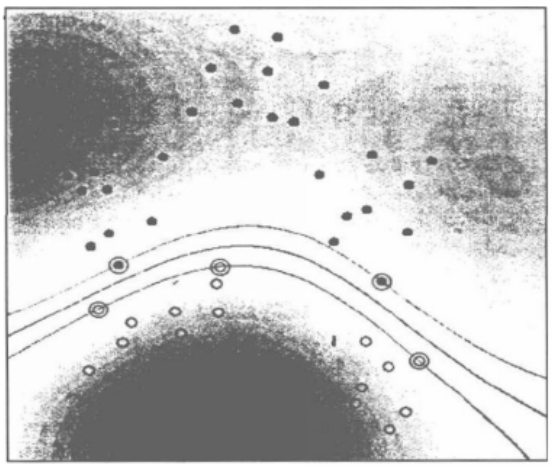


图 3.6　样本空间中的非线性划分

虽然在特征空间中似乎是解决了学习机的建立问题，但这里遗漏了一个重大的问题，那就是有关非线性映射 φ 的确定和计算，3.5 节的内容将解决这一问题。

3.4　Mercer 核和内积

我们试图通过引入核函数，从而绕过非线性映射的显式表达式来简化问题。首先讨论核与内积的关系。

Mercer 核和内积有着密切的关系。

回顾定理 3.1，Mercer 核可以展开成式(3.2)

$$K(\boldsymbol{X},\boldsymbol{Y}) = \sum_{i=1}\lambda_i\psi_i(\boldsymbol{X})\psi_i(\boldsymbol{Y}),$$

再回顾 3.2 节的非线性映射式(3.3)

$$\boldsymbol{X} \mapsto \varphi(\boldsymbol{X}) = (\sqrt{\lambda_1}\psi_1(\boldsymbol{X}),\ \sqrt{\lambda_2}\psi_2(\boldsymbol{X}),\cdots,\sqrt{\lambda_k}\psi_k(\boldsymbol{X}),\cdots),$$

如果我们定义特征空间上的内积为通常的欧几里得空间的内积(对应分量相乘再求和)，显然可以得到重要的关系式

$$K(\boldsymbol{X},\boldsymbol{Y}) = (\varphi(\boldsymbol{X})\cdot\varphi(\boldsymbol{Y}))。$$

该式说明一个 Mercer 核函数可以表示成内积的形式。反之，由任意内积定义的核是否是 Mercer 核呢？

可以验证：由内积定义的核一定是实对称的，进一步可以证明它是正定的，从而

它是 Mercer 核。这样我们可以把 Mercer 核的定义推广为：

定义 3.9 称由内积定义的核

$$K(\boldsymbol{X},\boldsymbol{Y}) = (\varphi(\boldsymbol{X}) \cdot \varphi(\boldsymbol{Y})) \tag{3.7}$$

为一个 Mercer 核。

同一个核可能由不同的内积或不同的非线性映射 φ 来定义。比如前面提到的 Mercer 核还可以如下定义：

例 3.5 设 Mercer 核 $K(\boldsymbol{X},\boldsymbol{Y})$的特征值和特征函数分别为 λ_i 和 $\psi_i(\boldsymbol{X})$，定义非线性映射 φ 为

$$\boldsymbol{X} \mapsto \varphi(\boldsymbol{X}) = (\varphi_1(\boldsymbol{X}), \cdots, \varphi_k(\boldsymbol{X}), \cdots) = (\lambda_1\psi_1(\boldsymbol{X}), \lambda_2\psi_2(\boldsymbol{X}), \cdots, \lambda_k\psi_k(\boldsymbol{X}), \cdots),$$

定义特征空间中的内积为$(\varphi(\boldsymbol{X}) \cdot \varphi(\boldsymbol{Y})) = \sum_i \frac{\varphi_i(\boldsymbol{X})\varphi_i(\boldsymbol{Y})}{\lambda_i}$，则很容易验证式(3.7)也成立。通常称由 Mercer 核如此定义了映射 φ 和内积而生成的特征空间为一个再造核希尔伯特空间(Reproducing Kernel Hilbert Space，RKHS)。

如果通过已知非线性映射 φ 来求核函数 $K(\boldsymbol{X},\boldsymbol{Y})$，则依据式(3.7)很容易就能算出，如例 3.6。

例 3.6 设 $\boldsymbol{X}=(x_1,x_2,x_3)$，$\boldsymbol{Y}=(y_1,y_2,y_3)$，非线性映射 φ 为

$$\varphi(\boldsymbol{X}) = (x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, 1),$$

$$\varphi(\boldsymbol{Y}) = (y_1^2, y_2^2, y_3^2, \sqrt{2}y_1y_2, \sqrt{2}y_1y_3, \sqrt{2}y_2y_3, \sqrt{2}y_1, \sqrt{2}y_2, \sqrt{2}y_3, 1),$$

则
$$\begin{aligned}(\varphi(\boldsymbol{X}) \cdot \varphi(\boldsymbol{Y})) &= (x_1^2y_1^2 + x_2^2y_2^2 + x_3^2y_3^2 + 2x_1x_2y_1y_2 + 2x_1x_3y_1y_3 \\ &\quad + 2x_2x_3y_2y_3 + 2x_1y_1 + 2x_2y_2 + 2x_3y_3 + 1) \\ &= (x_1y_1 + x_2y_2 + x_3y_3 + 1)^2 \\ &= (\boldsymbol{X} \cdot \boldsymbol{Y} + 1)^2 \\ &= K(\boldsymbol{X},\boldsymbol{Y})。\end{aligned}$$

但若通过已知核函数 $K(\boldsymbol{X},\boldsymbol{Y})$来求非线性映射 φ 和内积，则很不容易，例 3.7 是一个简单例子。

例 3.7 设 $\boldsymbol{X}=(x_1,x_2)$，$\boldsymbol{Y}=(y_1,y_2)$，核函数为 $K(\boldsymbol{X},\boldsymbol{Y}) = (\boldsymbol{X} \cdot \boldsymbol{Y}+2)^2$，求线性映射 φ。

由
$$\begin{aligned}K(\boldsymbol{X},\boldsymbol{Y}) &= (\boldsymbol{X} \cdot \boldsymbol{Y} + 2)^2 = (x_1y_1 + x_2y_2 + 2)^2 \\ &= (x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 + 4x_1y_1 + 4x_2y_2 + 4)\end{aligned}$$

$$= ((x_1, x_2, \sqrt{2}x_1x_2, 2x_1, 2x_2, 2) \cdot (y_1, y_2, \sqrt{2}y_1y_2, 2y_1, 2y_2, 2))$$

$$= (\varphi(\boldsymbol{X}) \cdot \varphi(\boldsymbol{Y})),$$

从而看出

$$\boldsymbol{X} \mapsto \varphi(\boldsymbol{X}) = (x_1, x_2, \sqrt{2}x_1x_2, 2x_1, 2x_2, 2)。$$

这里其实就是“凑”内积的办法。如果对于一般的高次多项式核 $K(\boldsymbol{X},\boldsymbol{Y}) = (a\boldsymbol{X} \cdot \boldsymbol{Y} + c)^d$，这种办法实在太“笨拙”了。而如果对于其他类的核，我们简直连这种“笨拙”办法也没有。幸运的是在 SVM 方法中我们并不需要真正求出这个非线性映射 φ 的显式表达式来。

Mercer 核很多，比较常见的 Mercer 核如下（其中 a, c, d, g, u 均为参数）：

（1）多项式核

$$K(\boldsymbol{X},\boldsymbol{Y}) = (a\boldsymbol{X} \cdot \boldsymbol{Y} + c)^d;$$

（2）高斯函数核（也称作径向基核 Radial Basis Function，RBF）

$$K(\boldsymbol{X},\boldsymbol{Y}) = e^{-g\sum_{i=1}^{n}(x_i - y_i)^2};$$

（3）柯西函数核

$$K(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{1}{1 + u(x_i - y_i)^2};$$

（4）拉普拉斯函数核

$$K(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n} e^{-u|x_i - y_i|};$$

（5）双曲正割函数核

$$K(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{2}{e^{u(x_i - y_i)} + e^{-u(x_i - y_i)}};$$

（6）平方正弦函数核

$$K(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{\sin^2[u(x_i - y_i)]}{u^2(x_i - y_i)^2} \qquad (u > 0)。$$

此外，由已知的 Mercer 核经过某些运算可以生成新的 Mercer 核，下面列出一些常见的构造方法。

设 $K_1(\boldsymbol{X},\boldsymbol{Y})$，$K_2(\boldsymbol{X},\boldsymbol{Y})$ 为 Mercer 核，$\boldsymbol{B}$ 为实对称正定阵，f 为实函数，α 为实数，$\boldsymbol{P}$ 为具有正系数的多项式，φ 为 $\boldsymbol{X} \mapsto \boldsymbol{R}^m$ 的映射，则下面列出的 $K(\boldsymbol{X},\boldsymbol{Y})$ 都是 Mercer 核

$$K(\boldsymbol{X},\boldsymbol{Y}) = K_1(\boldsymbol{X},\boldsymbol{Y}) + K_2(\boldsymbol{X},\boldsymbol{Y}),$$

$$K(\boldsymbol{X},\boldsymbol{Y}) = \alpha K_1(\boldsymbol{X},\boldsymbol{Y}),$$

$$K(\boldsymbol{X},\boldsymbol{Y}) = f(\boldsymbol{X})f(\boldsymbol{Y}),$$

$$K(\boldsymbol{X},\boldsymbol{Y}) = K_1(\varphi(\boldsymbol{X}),\varphi(\boldsymbol{Y})),$$

$$K(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{X}'\boldsymbol{B}\boldsymbol{Y},$$

$$K(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{P}(K_1(\boldsymbol{X},\boldsymbol{Y})),$$

$$K(\boldsymbol{X},\boldsymbol{Y}) = \exp(K_1(\boldsymbol{X},\boldsymbol{Y}))。$$

3.5 基于核方法的非线性 SVM

我们已经做好了必要的准备，现在回到特征空间的非线性分类问题。

在 3.3 节，我们把特征空间上的线性可分离的问题归结为式(3.5)

$$\max J(\boldsymbol{\alpha}) = \max\left\{\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{\substack{i=1\\j=1}}^{l}\alpha_i\alpha_j\, y_i y_j(\varphi(\boldsymbol{X}_i)\cdot\varphi(\boldsymbol{X}_j))\right\},$$

$$\text{约束条件：}\sum_{i=1}^{l}\alpha_i y_i = 0 \qquad (\alpha_i \geqslant 0; i = 1,\cdots,l)。$$

最后求得最优超平面决策函数为式(3.6)

$$M(\boldsymbol{X}) = \operatorname{sgn}\Big(\sum_{\text{支持向量}}\alpha_i^*\, y_i(\varphi(\boldsymbol{X})\cdot(\varphi(\boldsymbol{X}_i)) + b^*\Big)$$

式(3.5)和(3.6)中均出现了样本空间的点积$(\varphi(\boldsymbol{X}_i)\cdot\varphi(\boldsymbol{X}_j))$，如果我们能够直接“整体”算出点积式来，则没有必要得到难求的非线性映射 φ 了。这是完全办得到的：

由于内积实际上是一个 Mercer 核(据定义 3.8)，将式(3.7)

$$K(\boldsymbol{X},\boldsymbol{Y}) = (\varphi(\boldsymbol{X})\cdot\varphi(\boldsymbol{Y}))$$

代入式(3.5)和(3.6)最后得到

$$\max J(\boldsymbol{\alpha}) = \max\left\{\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{\substack{i=1\\j=1}}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{X}_i,\boldsymbol{X}_j)\right\} \tag{3.8}$$

$$\text{约束条件：}\sum_{i=1}^{l}\alpha_i y_i = 0 \qquad (\alpha_i \geqslant 0; i = 1,\cdots,l)$$

$$M(\boldsymbol{X}) = \operatorname{sgn}\Big(\sum_{\text{支持向量}}\alpha_i^*\, y_i K(\boldsymbol{X},\boldsymbol{X}_i) + b^*\Big) \tag{3.9}$$

式(3.8)是在特征空间中实际上需要求解的最优化问题的表示式。式(3.9)是在特征空间中所求得的基于核方法的非线性 SVM 的最后表达式，也就是非线性支持向量学习机的最终分类决策函数。可以看出，虽然用到了特征空间及非线性映射，但实际计算中并不需要知道它们的显式表达。只需要求出支持向量 $\boldsymbol{X}_i$ 及其支持的"强度"α_i^* 和阈值 b^*，通过核函数的计算，即可得到原来样本空间的非线性划分输出值。

这样我们就通过核函数和线性 SVM 方法解决了非线性 SVM 问题。将线性 SVM 的算法归结为一个凸约束条件下的二次凸规划问题，对此已有很多成熟的算法和应用软件可资使用。对于特征空间中的线性不可分和多类划分问题，用核函数取代其中的内积运算，也可以同样得到解决，这里不再重述。至此，我们通过引入核函数，以及应用线性学习机的方法，彻底解决了原来样本空间中的高度非线性分类问题。

可以看出：当我们把样本空间通过非线性映射映入到高维甚至是无穷维的特征空间时，如果只运用了映象的点积，则可以用相对应的核函数来代替，而不需要知道映射的显式表达式。这是从线性 SVM 走向非线性 SVM 的关键一步。正因为如此，这种方法又被称为是基于核的 SVM 方法。这一思路还启发我们：在任何其他线性问题中，只要最后是归结为含有点积的表达式，都可以尝试用核函数进行替换，也许会有意想不到的结果。第 4 章的支持向量回归就是一例。此外，在运用判别分析、时间序列分析等方法时也可以借助这种核方法。

由于线性 SVM 可以看作是一种采用了特殊的核——线性核的 SVM，所以之后我们不再区别 SVM 是线性还是非线性的，而笼统地称之为 SVM 方法。

3.6　SVM 方法的特点

经过以上分析，我们再强调一下 SVM 方法的特点：

(1) SVM 方法是新颖的统计学习方法，它的建模过程与传统的统计方法完全不同，根本没有涉及概率测度，而是依据间隔最大化原则进行统计建模。

(2) 由于核函数的引入，应用 SVM 方法建立的学习机分类模型是非线性的模型，SVM 是一种处理非线性分类的有效方法。对于一般的分类学习问题，只要样本数据有效且足够多，应用 SVM 方法建立的模型的分类效果一般优于其他方法。

(3) 由于有较为严格的统计学习理论作保证，应用 SVM 方法建立的模型具有较好的推广能力，可以给出学习结果的推广能力和误差的界的估计，这是任何其他分类方法所不具备的。当然，目前这些界的估计还是很初步的。

(4) 从所建立的 SVM 模型可以看出：它只依赖于称为支持向量的少数样本，而且不同支持向量样本对分类结果所作贡献的大小是不同的，非支持向量的增减对建模没有影响。这不但可以帮助我们抓住关键样本，而且可以"剔除"大量冗余样本。这种依赖于关键样本建模的学习方法，与在不同因子间加权的人工神经网等方法具

有本质上的不同。当预报因子与预报对象间蕴涵的复杂非线性关系尚不清楚时，基于关键样本的方法可能优于基于因子的“加权”。

(5) SVM 方法具有较好的“鲁棒性”，这种鲁棒性表现在多个方面。首先是同一个样本集，增删非支持向量的样本对决策集没有影响；其次是对同一问题的不同样本集，所建模型有一定的稳定性；最后是对核函数，很多实验表明，SVM 方法对核的选取不敏感，同一问题应用不同核都可以得到接近的效果，不同核下建立的 SVM 往往包含同样的支持向量。

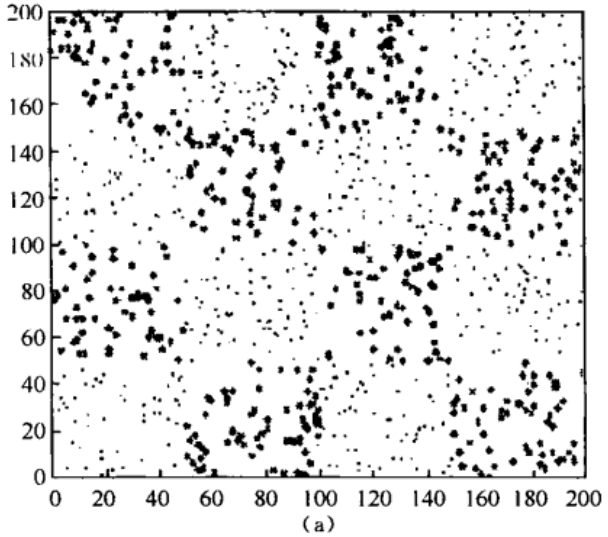
(6) SVM 方法的算法简单，易于实现，计算量与样本数成比例，与向量的维数无关。对于拥有几千个样本，每个样本是几百维的向量这样规模的问题，在我们的软件平台 CMSVM 上计算起来都很轻松。这也是其他学习方法所不能比拟的。

(7) SVM 的最终决策函数只由少数的支持向量所确定，计算的复杂性取决于支持向量的数目和核函数的计算，而不是样本空间的维数。由于内积与核函数的引入，SVM 方法的计算量与样本向量的维数无关，这在某种意义上避免了“维数灾”。

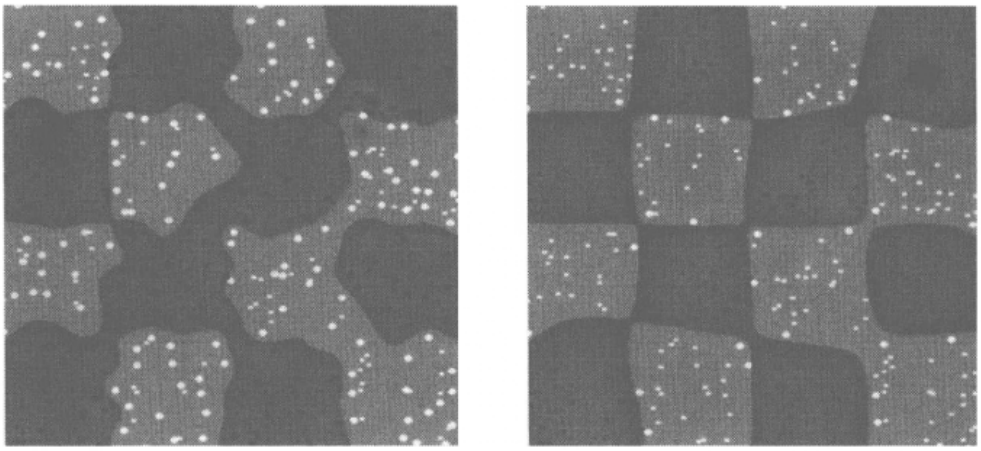
(8) 建立任何一个数据模型，人为的干预越少越客观。与其他方法相比，建立 SVM 模型所需要的先验干预较少，但核函数的选定及有关参数的优化仍是目前有待解决的问题之一，从例 3.8 可以看出这一点。

例 3.8 这是一个两类非线性划分的例子。图 3.7(a)为平面上待分类的两类数据，数据分布很像是国际象棋棋盘的一部分；图 3.7(b)和(c)为应用 SVM 方法由决策函数划分出来的两类区域，采用的是 RBF 核，左右两图应用的是同一核下的不同参数。两图中较大的白圆点是支持向量样本。

很明显，图 3.7(c)的分类效果较好，但它的支持向量反而较少(模型简单)，并且边界简单；图 3.7(b)的支持向量多、边界复杂，分类效果反而不好。这一结果虽不具备普遍性，因为此例具有局部线性的特性，但它至少说明应用 SVM 方法同样存在“过学习”和“欠学习”的问题，核参数的选取仍是一个有待解决的问题。



(a)



(b)　　(c)

图 3.7　SVM 分类示例

(a)为两类待分类样本点;(b)和(c)为选用不同核参数得到的 SVM 分类结果

3.7　本章小结

(1) 本章简单介绍了理解核方法所必需的数学知识,但是限于篇幅的原因,介绍是不系统、不严格和不全面的。

(2) 通过引入非线性映射,把建模问题的样本空间映射到一个高维以至于无穷维空间中。在这个更为广阔的空间中采用了第 2 章介绍的线性 SVM 方法,从而建立起特征空间中的线性 SVM 模型。

(3) Mercer 定理的应用是从线性学习机过渡到非线性学习机的关键。由于核函数的引入,使我们无须知道所引入的非线性映射的具体表达式,就建立起了样本空间上的非线性 SVM 模型,而且没有增加算法的复杂度。

(4) SVM 方法具有很多优点,适用于解决非线性分类问题。

第4章　支持向量机回归分析

机器学习需要解决的另一个重要难题就是回归分析。回归分析又称函数估计，是多元统计分析的一个重要分支。它是比模式识别更一般、更困难的问题。它的基本思路是通过有限个样本数据，建立起反映输出和输入之间联系的函数关系，这种函数关系通常是连续的。如果问题在本质上是线性的，则求得的函数关系是线性函数，称为线性回归分析，对此已经有很多有效的方法和成功的应用，如天气预报中使用的模式输出统计(Model Output Statistic，MOS)预报方法。如果输出和输入之间的联系是非线性关系，则称为非线性回归分析，求出这种非线性关系的表达式或近似表达式变得十分困难，至今尚无完美的方法。

基于核函数的SVM方法给出了解决非线性回归问题的一种新思路，并且与已有的方法相比，具有很多特色。有趣的是，对于分类和回归两种不同的问题，却可以用非常类似的SVM方法得到解决。本章讨论的顺序和结构与第2和3章的SVM分类基本类似，即先定义最优回归超平面，进而解决线性SVM回归问题，最后引入核函数，解决非线性回归问题。

目前，人们已经提出了很多回归分析的方法。除非是为了进行比较对照，本书不涉及其他方法而只介绍SVM方法。Meyer D等(2002)把SVM方法与9种已有的回归分析方法作了较为系统的比较分析，认为SVM是一种有效的回归分析方法。

4.1　回归分析的问题表述

回归分析又称函数估计，它要解决的问题是：依据有限的观测数据(训练样本)来寻求蕴涵着的回归函数，进而用求得的回归函数对未来数据(预报数据)进行预报。这类回归分析问题可以进行如下数学描述：

给定一组训练样本集

$$(\boldsymbol{X}_1, y_1), (\boldsymbol{X}_2, y_2), \cdots, (\boldsymbol{X}_l, y_l) \tag{4.1}$$

其中 $\boldsymbol{X}_i \in \boldsymbol{R}^n$，为 n 维向量，通常为预报因子值；$y_i \in \boldsymbol{R}$，为预报对象值；

给出待预报样本的预报因子数据集

$$\boldsymbol{X}_{l+1}, \boldsymbol{X}_{l+2}, \cdots, \boldsymbol{X}_{l+m} \tag{4.2}$$

寻求与训练样本的输入输出拟合最优的函数关系 $y = f(\boldsymbol{X})$，进而求出预报对象 y_i 的输出值。

上述问题当 $n=1$ 时称为一元回归分析，当 $n>1$ 时称为多元回归分析。如果所

得函数关系 $y = f(x)$是线性函数，则称为线性回归，否则为非线性回归。图 4.1 给出二维数据的线性与非线性回归图示。

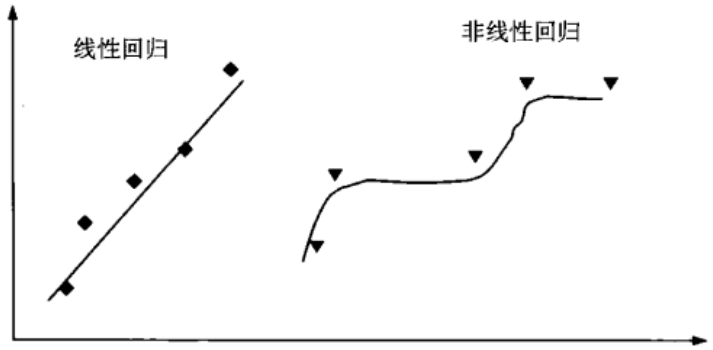
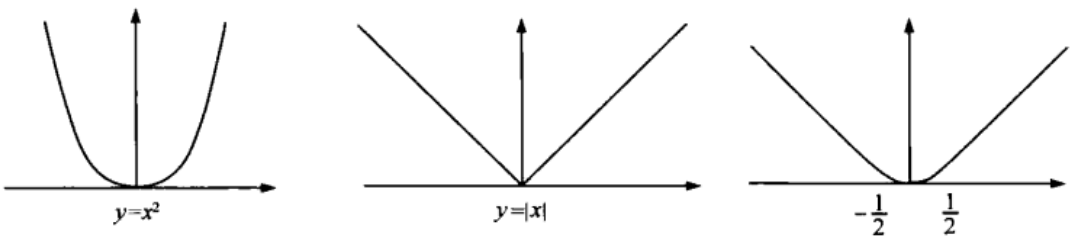


图 4.1　线性与非线性回归图示

4.2　ε-不敏感函数

回归分析中首先需要考虑的是所建立的回归模型的误差计算，即回归效果的估计。相比之下，在模式识别中这一问题比较简单，那里只需要累加计算错误分类的样本个数。但在回归分析中，要想在指定的函数类中选取出“拟合最优”的函数关系 $y = f(\boldsymbol{X})$，必须明确评优的标准。

这里的“拟合最优”是指按某一规定的误差函数来计算，所求得函数关系 $y = f(\boldsymbol{X})$对样本数据集拟合得“最好”，即所有训练样本的累计误差为最小。有多种计算累计误差的算法，图 4.2 给出了多元统计分析中常用的三种误差函数。



(a)平方函数　　(b)绝对值函数　　(c)Huber 函数

图 4.2　常用误差函数

(图中横坐标为实际误差，纵坐标为参与累计的误差值)

图 4.2(a)为平方误差函数，它对应于通常计算的均方误差或均方根误差，其特点是对于小误差不敏感对大误差反映敏感；图 4.2(b)为通常的绝对误差，它对大小误差敏感度一样；图 4.2(c)为图 4.2(a)和(b)的结合($|x|<\frac{1}{2}$为图 4.2(a)，其余为 $|x|-\frac{1}{2}$)，它保持了图 4.2(a)的对于小误差不敏感的优点，克服了对大误差过于敏感的弱点。

为了讨论 SVM 回归，Vapnik 定义了 ε-不敏感误差函数(如图 4.3)。

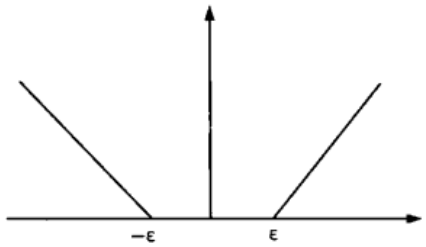


图 4.3 ε-不敏感误差函数

其数学表达式为

$$L_\varepsilon(y)=\begin{cases}0 & \text{当}\ |f(x)-y|<\varepsilon\\ |f(x)-y|-\varepsilon & \text{其他}\end{cases} \tag{4.3}$$

ε-不敏感误差函数中的 ε 为参数，当 ε= 0 时，式(4.3)等同于图 4.2(b)。当 ε≠0 时，式(4.3)相当于把图 4.2(b)中的两条射线分别向左右平移 ε 距离变成图 4.3。

其含义为：当误差小于 ε 时，误差忽略不计；当误差超过 ε 时，误差函数的值为实际误差减去 ε。换言之，这种误差函数中间有一个宽度为 2ε 的不敏感带，以后我们称其为 ε 带或 ε 管道。

以下对 SVM 回归的讨论就是在假设这种误差函数存在的条件下进行的。

4.3 最优回归超平面与 SVM 线性回归

前两章介绍的 SVM 方法可以类似地应用于回归分析。回归分析中最简单的是线性回归，从而我们首先讨论 SVM 的线性回归方法。

4.3.1 最优回归超平面

与普通线性回归问题一样，我们的目的是寻求一个回归超平面，依据 ε-不敏感误差函数来度量误差，使它对训练样本数据集拟合得最好，我们称这样的超平面为最优回归超平面。

与 SVM 分类问题所不同的是：回归学习的样本点只有一类而不是两类(或多类)，所寻求的最优回归超平面不是使两类样本点分得“最开”，而是使所有样本点离超平面的“总偏差”最小。这时样本点都在两条边界线之间，求最优回归超平面同样等价于求最大间隔，推导过程与 SVM 分类情况相同，这里略去。图 4.4 给出最优回归超平面的粗略图示。为了方便理解，这里采用了 SVM 分类最优划分超平面示意图(图 4.4)的同一框架。

由此看出，与 SVM 分类的情况相同，最优回归超平面只由落在边界上的样本点就可以完全确定，如图 4.4 中的 x_1，x_2 等。这些样本点我们同样称为支持向量。非

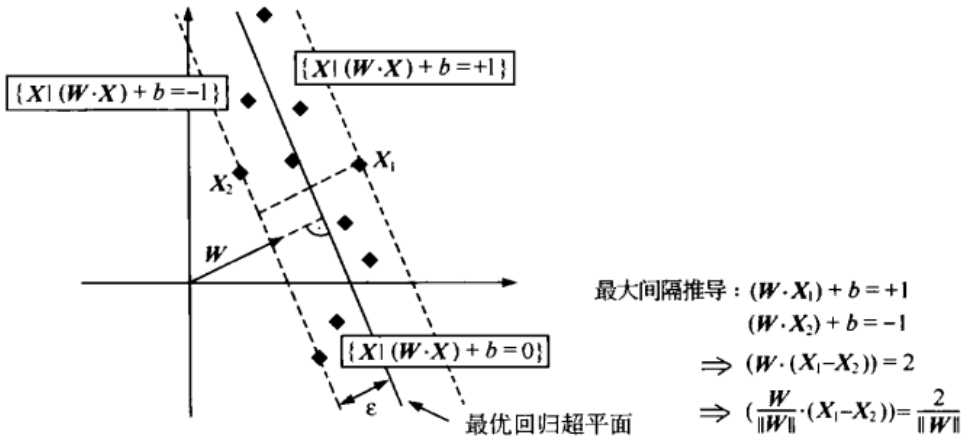


图 4.4　最优回归超平面图示

支持向量样本点对于建立最优回归超平面没有贡献。

4.3.2　支持向量机线性回归

如果采用图 4.3 的 ε-不敏感函数作为误差函数，首先考虑总偏差为零的情况，即所有样本点都可以落入 ε 管道中，这相当于 SVM 分类时的线性可分的情况。此时所有样本点到所求超平面的距离都小于 ε，采用 2.3 节同样的分析方法，寻求最优回归超平面的问题转化为求解如下一个二次凸规划问题

$$\min\left\{\frac{1}{2}\|\boldsymbol{W}\|^2\right\}$$

$$\text{约束条件：}y_i-(\boldsymbol{W}\cdot\boldsymbol{X}_i)-b\leqslant\varepsilon \tag{4.4}$$

$$(\boldsymbol{W}\cdot\boldsymbol{X}_i)+b-y_i\leqslant\varepsilon \qquad (i=1,\cdots,l)$$

式(4.4)与第 2 章的式(2.7)是同一类最优化问题，因此可以用同样方法求得最优回归超平面的表达式。

实际问题中往往有噪音干扰，不可能所有样本点都落在 ε 管道中。即使一时是这样，随着 ε 的减小，也会有个别样本点“溢出”。当个别样本点到所求超平面的距离大于 ε 时，这相当于 SVM 分类时的线性不可分情况，ε-不敏感函数使超出的偏差相当于 SVM 分类中引入的松弛变量 ξ_i(图 4.5)。

通过修改式(4.4)的目标函数(考虑落到 ε 管道之外的总偏差)和约束条件，有偏离点的最优回归超平面归结为一个新的二次凸规划问题

$$\min\left\{\frac{1}{2}\|\boldsymbol{W}\|^2+\boldsymbol{C}\sum_{i=1}^{l}(\xi_i+\xi_i^*)\right\}$$

$$\text{约束条件：}y_i-(\boldsymbol{W}\cdot\boldsymbol{X}_i)-b\leqslant\varepsilon+\xi_i \tag{4.5}$$

$$(\boldsymbol{W}\cdot\boldsymbol{X}_i)+b-y_i\leqslant\varepsilon+\xi_i^* \qquad (\xi_i\geqslant 0,\xi_i^*\geqslant 0)$$

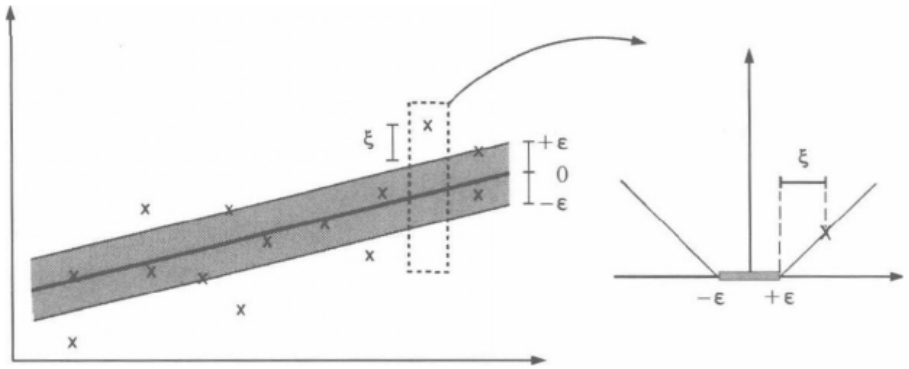


图 4.5 有偏离点的最优回归超平面示意图

其中 ξ_i 与 ξ_i^* 分别对应于样本点在最优回归超平面上方和下方的两种情形；C 为事先给定的惩罚系数。

对于最优化问题式(4.4)和(4.5)，由于目标函数和约束条件都是凸集，根据最优化理论，它们都存在唯一的全局最小解。

4.3.3 求解最优回归超平面

与 SVM 分类最优超平面的求法类似，采用 Lagrange 乘数法并应用 KKT 条件，可求得式(4.4)和(4.5)的最优回归超平面。若所有的 ξ_i 与 ξ_i^* 均为零，则式(4.5)即为式(4.4)。因此，下面仅以式(4.5)的求解过程为例进行讨论。

首先定义 Lagrange 函数如下

$$L(\boldsymbol{W},b,\boldsymbol{\alpha},\varepsilon,\boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{W}\|^2 + C\sum_{i=1}^{l}(\xi_i+\xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon+\xi_i-y_i+(\boldsymbol{W}\cdot\boldsymbol{X}_i)+b) -$$

$$\sum_{i=1}^{l}\alpha_i^*(\varepsilon+\xi_i^*+y_i-(\boldsymbol{W}\cdot\boldsymbol{X}_i)-b) - \sum_{i=1}^{l}(\eta_i\xi_i+\eta_i^*\xi_i^*) \tag{4.6}$$

对 $\boldsymbol{W}$，b，ξ_i，ξ_i^* 求偏导数并令偏导数为零，得到

$$\partial_b L = (\alpha_i^* - \alpha_i) = 0 \tag{4.7}$$

$$\partial_w L = \boldsymbol{W} - (\alpha_i - \alpha_i^*)\boldsymbol{X}_i = 0 \tag{4.8}$$

$$\partial_\xi L = C - \alpha_i^* - \eta_i^* = 0 \tag{4.9}$$

将这三式带入式(4.6)，可以消去 $\boldsymbol{W}$，b，ξ_i，ξ_i^*，η_i，η_i^* 得到式(4.5)的对偶最优化公式

$$\max J(\boldsymbol{\alpha}) = \max\left\{-\frac{1}{2}\sum_{\substack{i=1\\j=1}}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(\boldsymbol{X}_i\cdot\boldsymbol{X}_j) - \varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*)\right\}$$

$$\text{约束条件：}\sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0 \tag{4.10}$$

$$0 \leqslant \alpha_i,\ \alpha_i^* \leqslant C \qquad (i = 1,\cdots,l)$$

式(4.8)即转化为

$$\boldsymbol{W} = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\boldsymbol{X}_i \tag{4.11}$$

通过求解式(4.10)可以得到 α_i 和 α_i^*，从而由式(4.11)计算得到 $\boldsymbol{W}$，而为了求出系数 b，要用到 KKT 条件

$$\alpha_i(\varepsilon + \xi_i - y_i + (\boldsymbol{W}\cdot\boldsymbol{X}_i) + b) = 0,$$

$$\alpha_i^*(\varepsilon + \xi_i^* + y_i - (\boldsymbol{W}\cdot\boldsymbol{X}_i) - b) = 0,$$

$$(C - \alpha_i)\xi_i = 0,$$

$$(C - \alpha_i^*)\xi_i^* = 0。$$

从这些条件中可以得出如下结果：

(1) 只有对应于 $\alpha_i = C$ 或 $\alpha_i^* = C$ 的样本$(\boldsymbol{X}_i, y_i)$会留在 ε 管道之外；

(2) $\alpha_i\alpha_i^* = 0$，这说明任何样本$(\boldsymbol{X}_i, y_i)$在两个方向上的松弛变量不能同时非零；

(3) 当 $0 < \alpha_i < C$ 时，$\xi_i = 0$；当 $0 < \alpha_i^* < C$ 时，$\xi_i^* = 0$，此时对应的支持向量落在边界上；

(4) 落在管道内的样本点(非支持向量)对应的 α_i 和 α_i^* 同时为零，所以在最终最优回归超平面解析表达式中不出现这些样本点；

(5) 可以很容易通过边界上的支持向量样本点，由式(4.12)的两式计算出系数 b

$$\begin{aligned} b &= y_i - (\boldsymbol{W}\cdot\boldsymbol{X}_i) - \varepsilon \qquad (\text{当 } \boldsymbol{X}_i \text{ 属于上边界，即 } 0 < \alpha_i < C) \\ b &= y_i - (\boldsymbol{W}\cdot\boldsymbol{X}_i) + \varepsilon \qquad (\text{当 } \boldsymbol{X}_i \text{ 属于下边界，即 } 0 < \alpha_i^* < C) \end{aligned} \tag{4.12}$$

图 4.6 给出上述关系的直观描述，$\boldsymbol{X}_1$，$\boldsymbol{X}_2$ 和 $\boldsymbol{X}_3$ 对应于 $\xi_i = 0$ 而 $0 < \alpha_i < C$ 的情形；$\boldsymbol{X}_4$ 和 $\boldsymbol{X}_5$ 对应于 $\xi_i \neq 0$ 而 $\alpha_i = C$；其余样本点均落在 ε 管道内，则 ξ_i 和 α_i 均为零。最后求得的最优线性回归超平面的解析表达式为

$$f(\boldsymbol{X}) = (\boldsymbol{W}\cdot\boldsymbol{X}) + b = \sum_{\text{支持向量}}(\alpha_i - \alpha_i^*)(\boldsymbol{X}\cdot\boldsymbol{X}_i) + b \tag{4.13}$$

在上述求解过程中，有多种高效算法和成熟的计算机程序可资利用。最后结果表明：最优回归超平面只由作为支持向量的样本点完全确定。

与分类问题类似，线性回归也是早已基本解决的问题，我们之所以进行上述讨论，完全是为了引出非线性回归问题的解决方法。式(4.13)中出现的点积提示我们

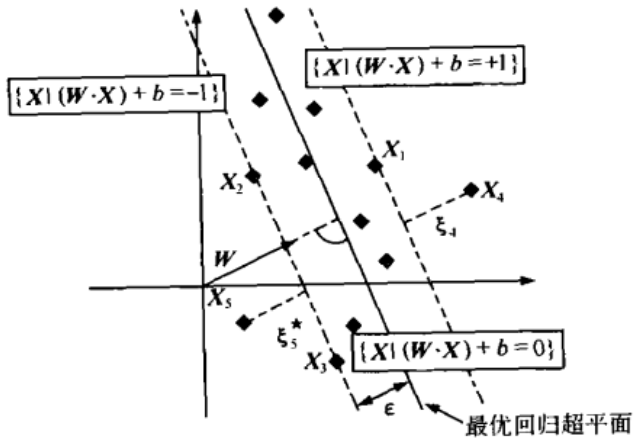


图 4.6 样本点及其对应的 α,ξ 的关系

完全可以通过核函数来引入非线性映射，在特征空间中作上述 SVM 的线性回归，从而实现了样本空间的高度非线性回归。

4.4 非线性 SVM 回归

为了解决本质上非线性的回归问题，同样采用引入非线性映射 φ 的方法，把样本空间$\mathscr{X}$映射入一个高维(以至于无穷维)的特征空间$\mathscr{F}$(Hilbert 空间)，从而使在样本空间中的高度非线性问题在特征空间中可以通过应用 SVM 线性回归方法得以解决。图 4.7 给出这一思路的直观说明。

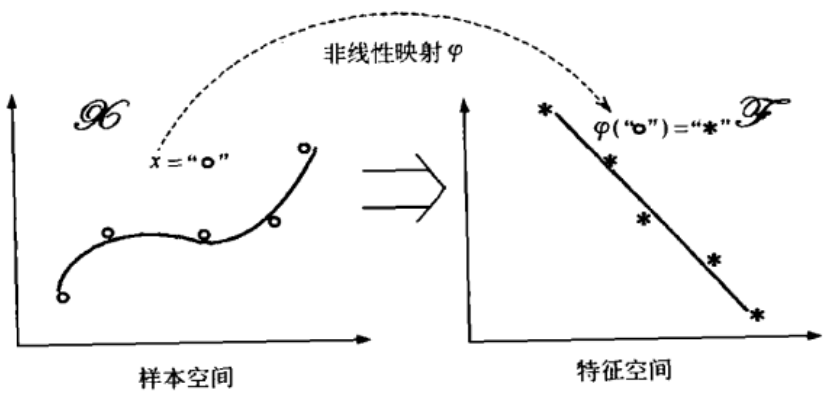


图 4.7 回归样本空间到特征空间的非线性映射

由于引入了非线性映射 φ，在特征空间$\mathscr{F}$中用 $\varphi(\boldsymbol{X}_i)$代替 $\boldsymbol{X}_i$，将求线性最优回归超平面的式(4.10)变为

$$\max J(\boldsymbol{\alpha})=\max\left\{-\frac{1}{2}\sum_{\substack{i=1\\j=1}}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(\varphi(\boldsymbol{X}_i)\cdot\varphi(\boldsymbol{X}_j))-\varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*)\right\}$$

$$\text{约束条件：} \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0 \tag{4.14}$$

$$0 \leqslant \alpha_i,\ \alpha_i^* \leqslant C \qquad (i = 1,\cdots,l)$$

求得特征空间中的最优线性回归超平面的解析表达式为

$$f(\boldsymbol{X}) = (\boldsymbol{W} \cdot \boldsymbol{X}) + b = \sum_{\text{支持向量}}(\alpha_i - \alpha_i^*)(\varphi(\boldsymbol{X}_i) \cdot \varphi(\boldsymbol{X}_j)) + b \tag{4.15}$$

由于内积实际上是一个 Mercer 核(见定义 3.9),将式(3.7)

$$K(\boldsymbol{X},\boldsymbol{Y}) = (\varphi(\boldsymbol{X}) \cdot \varphi(\boldsymbol{Y}))$$

代入式(4.14)和(4.15),就消去了非线性映射的显式表达式,最后得到

$$f(\boldsymbol{X}) = (\boldsymbol{W} \cdot \boldsymbol{X}) + b = \sum_{\text{支持向量}}(\alpha_i - \alpha_i^*)K(\boldsymbol{X},\boldsymbol{X}_i) + b \tag{4.16}$$

这就是在特征空间中所求得的基于核方法的非线性 SVM 回归模型的最后表达式,也就是非线性 SVM 的最优回归函数。虽然推导过程中运用到了特征空间及非线性映射,但实际计算中并不需要知道它们的显式表达。只需要求出支持向量对应的系数 α_i,α_i^* 和 b 的值,通过核函数的计算,即可得到原来样本空间的非线性回归函数。

这样我们就通过核函数和线性 SVM 回归方法解决了非线性 SVM 回归问题。这里再次强调:式(4.16)给出的是特征空间上的线性最优回归超平面,但与之相对应的却是样本空间上的非线性回归曲面,如图 4.8 所示。

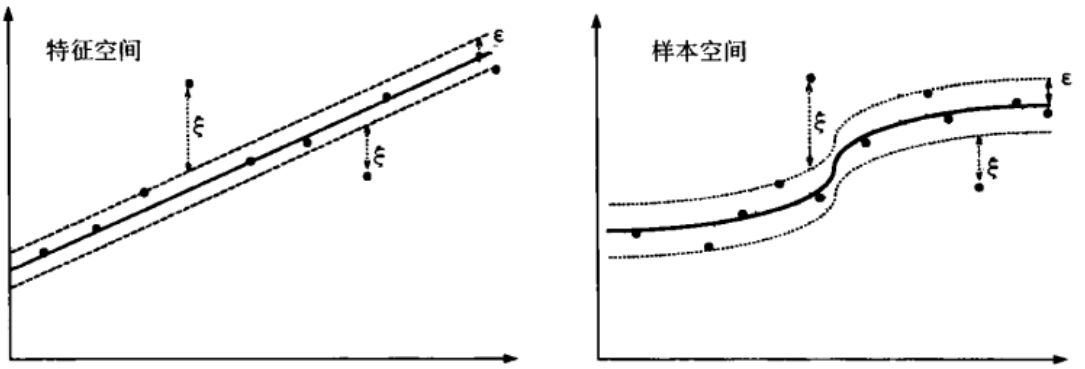


图 4.8　特征空间与样本空间的回归函数

由于线性 SVM 回归可以看作是一种采用了特殊的核——线性核的支持向量机,所以以后也不再区别线性还是非线性 SVM 回归,而笼统地称为 SVM 回归方法,或与 SVM 分类方法一起,简称为 SVM 方法。

4.5 SVM 回归方法的特点

与已有的其他回归分析方法相比，SVM 方法看起来似乎是有些怪异。这主要是因为它引入了 ε-不敏感误差函数和最优回归超平面。这一新颖的回归分析方法具有如下特点：

(1) 由于引入了样本空间到特征空间的非线性映射，所以 SVM 回归方法特别适用于处理本质上为非线性函数估计问题。对于采用已有的回归方法却因算法的复杂性而难于或不能解决的问题，不妨试用一下这一新方法。

(2) 由于引入了核函数而避开了非线性映射的显式表达式，其计算量与线性回归方法相当。与已有的其他非线性回归方法相比，SVM 回归算法简单、计算量小、易于实现，可以解决规模较大的问题。

(3) 对于通常的多元回归算法，计算的复杂性与样本点的维数密切相关。SVM 得到的回归函数只由少数的支持向量所确定，其复杂程度取决于支持向量的数目，而不是样本空间的维数，这在某种意义上避免了“维数灾”，大大扩展了多元回归的应用范围。通常几千个样本、每个样本维数为数百维的问题，计算起来毫无困难。

(4) SVM 回归方法具有较好的“鲁棒性”，这种鲁棒性表现在落入 ε 管道内的大量样本是否改变对最优决策函数毫无影响。

(5) 基于 ε-不敏感误差函数的 SVM 回归得到的最优回归函数，是在指定的函数类中，满足 ε 管道限制的“最平坦”的一个函数。ε 越大(允许偏差越大)，管道越“宽敞”，决定最优回归函数的支持向量越少，得到的回归函数越“平坦”。反之，ε 越小(允许偏差越小)，管道越“狭窄”，决定最优回归函数的支持向量个数越多，得到的回归函数越“曲折”。

下面通过一个仿真例子对上述特点加以直观说明。

例 4.1 假定待求的回归函数为 $f(x)=\dfrac{\sin x}{x}$，样本数据 (x_i, y_i) 是由 $y=f(x)$ 生成。图 4.9 给出 ε 管道与回归函数的“平坦性”关系的示意图。

图 4.9(a)为待求的回归函数，其余三个图分别为 ε 等于 0.1，0.2 和 0.5 时的 ε 管道及其对应的回归函数(图中的虚线)。显然，管道越宽，求得的回归函数越平坦。当然，误差也越大了。

图 4.10 从支持向量的角度直观解释 SVM 回归的特点。图 4.10 中的小黑圆点为样本点，大黑圆点为支持向量，实线为回归曲线。当 ε 管道宽度取不同值时，所要求的支持向量个数不同，随着 ε 的减小，支持向量的个数增加，同时回归曲线变得曲折但更接近客观最优值。图 4.10(a)中，ε 最大(0.5)，支持向量个数最少(3 个)，回归曲线最平坦，误差也最大；图 4.10(d)中，ε 最小(0.02)，支持向量个数最多(17 个)，回归曲线最曲折，但最接近客观最优函数，整体误差最小。

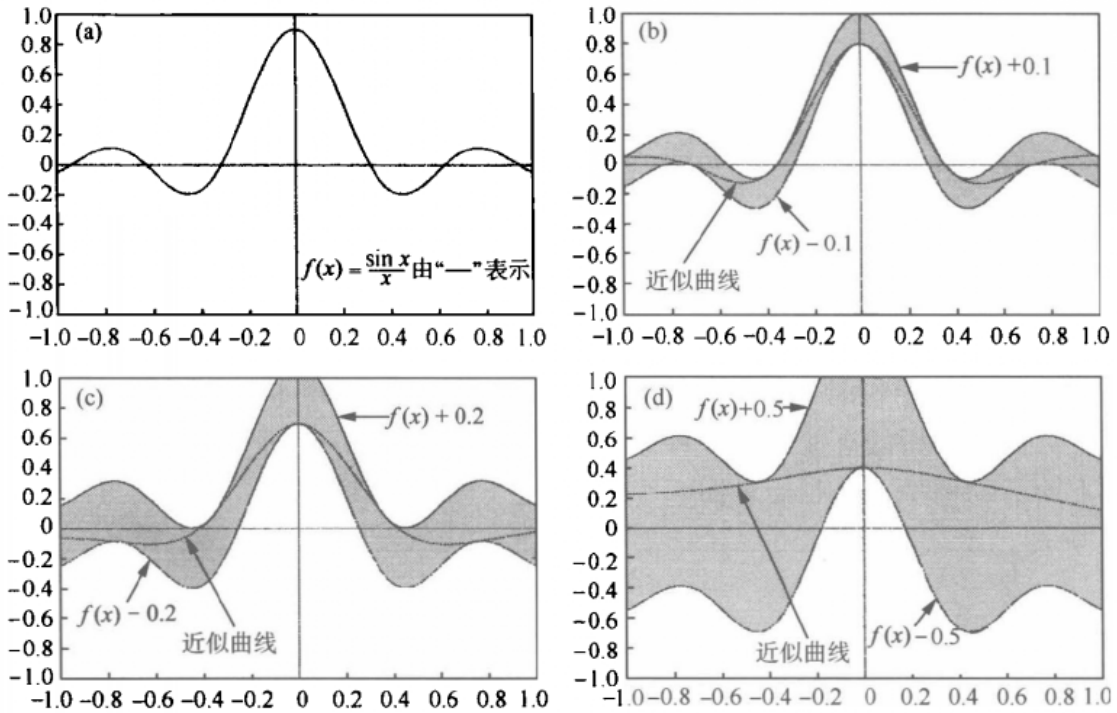


图 4.9　ε 管道与回归函数关系示意图

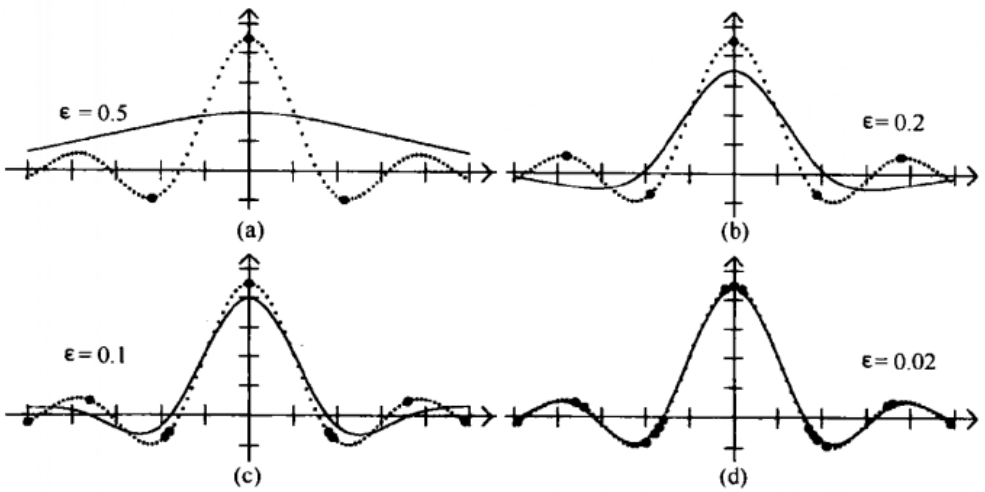


图 4.10　支持向量与回归函数关系示意图

（其中"·"代表样本点，"•"代表支持向量，"—"代表回归曲线）

图 4.9 和 4.10 提醒我们：

（1）在实际应用 SVM 回归方法时，一定要选取合适的 ε 管道精度，ε 太大和太小都不会导致好的结果。这里需要在先验误差和模型复杂性之间寻求一个折中，即 ε 太大则模型简单但误差大；ε 太小则误差小但模型复杂，甚至导致所谓的"过学习"。

（2）要满足训练样本的一定数量要求。如果问题很复杂而样本数又太少，即使全部样本都作为支持向量，也许还达不到精度的要求。当然，对于学习问题来说，一

般是训练样本越多越好。

(3) 这里似乎是验证了一个哲学的论断——单凭经验性的归纳决不能充分地证明必然性。在例 4.1 中,无论选取多么多的样本点,无论选取多么小的 ε,无论确定了多么多的支持向量,SVM 回归曲线可以多么无限靠近真实曲线,但它永远也不会变成 $f(x)=\frac{\sin x}{x}$(严格精确的真实曲线)。但从工程应用的角度看,这并不是问题,因为只要足够靠近真实就够用了。就连"真实曲线"本身,在实际应用时也只能取它的数值近似值。

最后给出 SVM 回归的一种物理上的解释。最优回归超平面同样满足 2.3 节关于最优划分超平面的有关物理解释,即它满足力和力矩的两个平衡条件,这里略去。

下面借助例 4.1 并且通过图 4.11 给出进一步的物理解释。

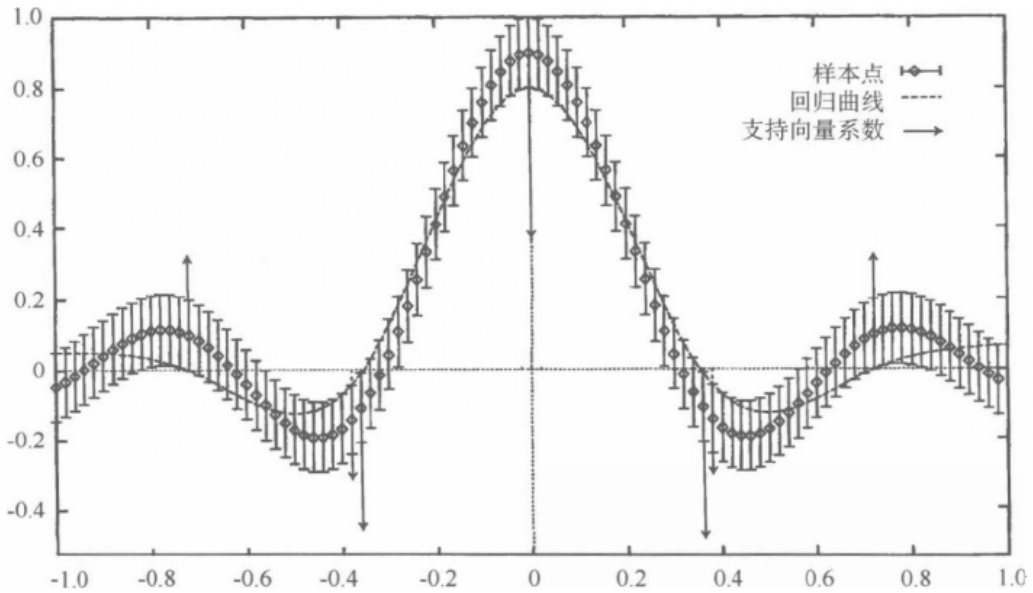


图 4.11 SVM 回归的物理解释

在图 4.11 中,ε 管道宽度为 0.2(ε= 0.1),图中出现 7 个支持向量,这些支持向量恰好都在 ε 管道边界上。图中的箭头表示的是对应支持向量的系数(α_i 或 α_i^*),其长短为系数的绝对值,方向指向管道内部。

可以作一个形象的比喻:最优回归曲线可以想象为是一条弹性极好的"钢丝",它恰好可以"插入"给定的 ε 管道中。"钢丝"与管道的接触点就是支持向量,管道壁对"钢丝"的支撑力使它在管道中达到平衡。当然,管道越细,"钢丝"与管道壁接触点越多,"钢丝"就变得越曲折。最优回归函数就是在管道确定的条件下,可以插入管道中的最为平坦的"钢丝"。

为什么支持向量样本一定要在 ε 管道壁上?这是由于支持向量对应的 α_i 或 α_i^* 不为零,由 KKT 条件可知支持向量样本一定要使式(4.5)约束条件取等号。这说明

这些样本一定在 ε 管道壁上。

还有一个值得提出的情况：当 ε 管道很宽使所有样本点都落在其中时，线性 SVM 回归总是选取管道的中轴作为最优回归函数。有时这是人们难以理解的，但也有其合理性如例 4.2 所示。

例 4.2　如图 4.12 所示，绝大多数样本点都分布在直线 $y=4$ 上，只有一个样本点落在直线 $y=2$ 上。如果做线性回归，直观上觉得若忽略个别点，回归直线应该就是直线 $y=4$；若兼顾到个别点时回归直线也应该在 $y=4$ 附近。

应用线性核做 SVM 回归计算得到的结果是：若 $\varepsilon \geqslant 1$，则回归曲线都是直线 $y=3$ 如图 4.12(a)，这一结果显然有些怪异；但当 $\varepsilon=0.01$ 时，回归曲线就恢复了正常[如图 4.12(b)]，回归直线为 $y=3.99$，实际上对于给定的 ε，此时 SVM 回归直线方程为 $y=4-\varepsilon$。通过此例可以看出，在实际应用中 ε 的确定也是至关重要的。

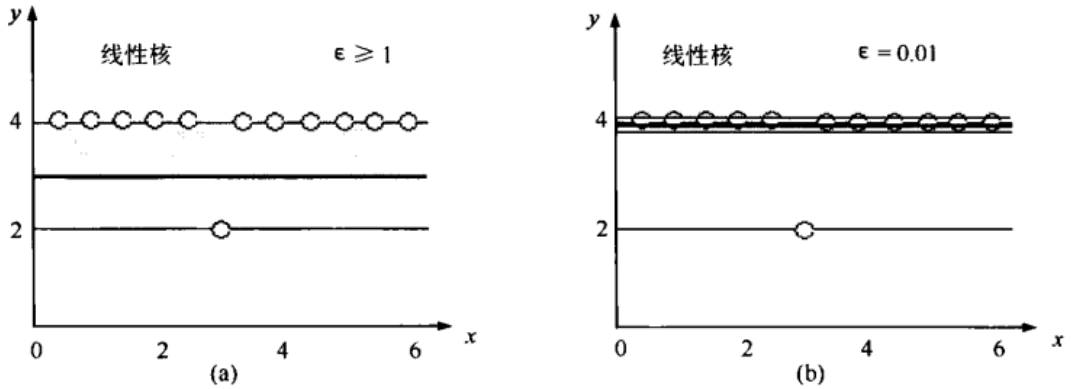


图 4.12　SVM 线性回归的例

如果把例 4.2 看作非线性问题来处理，应用径向基函数(RBF)核做 SVM 回归，回归曲线的图形大致如图 4.13 所示。

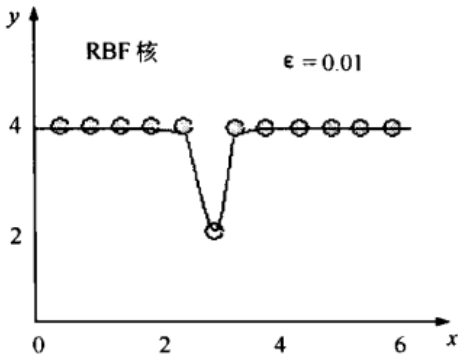


图 4.13　SVM 非线性回归的示例

应该指出，SVM 回归方法的参数选取仍是没有彻底解决的问题。除了 SVM 分类中遇到的核函数及核参数外，由于 ε-不敏感误差函数的引入，又多了一个 ε 参数(即回归带宽)，这些参数的选取直接影响回归效果，理想的参数优化方法还有待于人

们提出。还应该指出，同样由于 ε-不敏感误差函数的引入，采用常规回归效果的评估方法（如绝对差、均方差等）来对 SVM 回归方法的效果进行评估和比较，已经是不“公允”和不妥当的了。

必须强调指出：具有特色的 SVM 回归方法，由于引入了 ε-不敏感误差函数，大大减少了回归计算的运算量，有望取代最小二乘法，从而带来资料同化等方法的新进展。核方法这一先进思想对于新数值预报模式的建立也有重要价值。

4.6 本章小结

（1）本章首先引入 ε-不敏感误差函数，通过定义最优回归超平面和支持向量，把线性回归问题转化成一个凸约束条件下的凸规划问题，建立起线性 SVM 回归方法。

（2）进而通过引入非线性映射，把样本空间映射到一个特征空间，在特征空间中构建线性 SVM 模型。由于核函数的引入，使我们无须知道所引入的非线性映射的具体表达式，就建立起了样本空间上的非线性 SVM 回归模型。

（3）较为详细地分析了 SVM 回归方法的优点和弱点。

第5章　关于支持向量机方法的进一步思考

前面几章已经通过分类和回归的建模问题，初步介绍了 SVM 方法的基本思路。很显然，这是一种新颖的统计学习方法，它的基本框架具有一般性。在理解 SVM 基本方法的基础上，本章将从几个方面对这一方法进行深入一步的讨论。

5.1　从样本到样本的推理

经典逻辑中的主要推理方法为演绎推理和归纳推理。归纳是描述从特殊到一般的过程，演绎则描述从一般到特殊的过程。学习问题的通常解决途径为先从训练样本中归纳出一般的逼近函数——该函数反映了过程的内在关系，再由得到的一般函数演绎出测试样本点的值，即为了解决评估测试样本值的问题，先要建立更加普遍、一般和困难的逼近函数，图 5.1 给出这一过程的示意图。

SVM 方法采用的是如图 5.2 所示的解决模式：从给定训练样本出发直接导出未知函数在我们感兴趣的测试样本点的值，我们称之为直推模式。

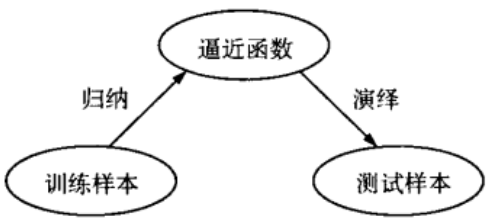


图 5.1　采用归纳—演绎推导过程的学习问题示意图

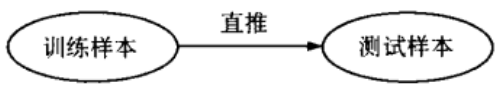


图 5.2　直推模式示意图

在解决一个给定的问题时，总是应该设法避免把解决一个更为一般的问题作为其一个中间步骤。在处理学习问题时，求测试点函数值这一问题与求出一个一般函数的问题相比，显然要来得简单。这就是说，传统的先归纳后演绎的推理方法，至少在处理机器学习时似乎是把问题“复杂化”了，即它试图通过解决一个困难得多的问题（求一般函数关系）来解决一个相对简单的问题（求一些测试点的值）。而 SVM 方法采用的是一种从样本到样本的直推捷径，不失为一种巧妙的做法①。

这种不同并不容易直接看出，因为 SVM 方法也同样要先求出学习机的决策函数来，再做预测。两者的区别在于：归纳—演绎推理最后求出的逼近函数通常是脱离样本的一般函数 $f(\boldsymbol{X})$，这个函数反映了过程的本质属性；SVM 方法得到的是由支

① 通常演绎推理过程为“个别→一般→个别”；而直推则是“个别→个别”，即不求一般（哲学上所谓的精确）。

持向量样本和核函数组合的决策函数 $M(\boldsymbol{X},K,\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m)$，由于它对核的不敏感性，SVM 的决策函数主要反映样本（特别是支持向量）的特性，并不直接代表实际过程的抽象属性。

5.2 SVM 方法的非线性本质

SVM 方法的应用重要性在于它可以较好地解决非线性问题。如何理解 SVM 方法的通用逼近能力和它的非线性本质？这里避开有关函数逼近的严格数学概念，只是结合实例加以直观说明。

机器学习问题的核心是“从给定的函数集 $f(\boldsymbol{X},\boldsymbol{\alpha})$ 中选择出能够最好逼近训练器（S）响应的函数”来。在 SVM 方法中，备选函数集的一般形式为

$$f(\boldsymbol{X},\boldsymbol{\alpha},\boldsymbol{X}_i)=\sum_{i=1}^{N}\alpha_i K(\boldsymbol{X},\boldsymbol{X}_i)+b \tag{5.1}$$

其中 K 为 Mercer 核，$N<l$ 为整数，$\boldsymbol{X},\boldsymbol{X}_i$ 为样本空间中的任意点。核 K 决定了学习机的类型。根据核的展开定理可以证明：当选取多项式、径向基函数或其他 Mercer 核（线性核除外）时，式（5.1）能以任意高的精度逼近任意给定的连续函数。而且，由于支持向量是在学习过程中自动确定的，所以式（5.1）中的项数 N 不需要事先确定。这说明 SVM 方法的备选函数集是非常“丰富”的，具有很好的通用逼近能力。

$f(\boldsymbol{X},\boldsymbol{\alpha},\boldsymbol{X}_i)$ 的通用逼近性说明了它的多样性，也注定了它的非线性性。式（5.1）给出的函数类相比我们较为熟悉的解析函数要复杂，它的多样性难以直观描述。下面给出极为简单的两个图示。

例 5.1 设核为一维径向基函数，样本向量的输入为一维。若只含有三个支持向量的学习机函数为

$$f(x)=0.3\exp(-(x-1)^2)-0.7\exp(-(x-3)^2)+0.4\exp(-(x+4)^2)-0.05$$

则其函数图形由 MATLAB 软件绘出（如图 5.3）。

从图 5.3 可以看出其非线性特性。随着支持向量个数的增加，其表现能力及非线性特性将更加明显。

图 5.4 给出某径向基函数核、样本向量输入为二维且含有六个支持向量的学习机函数的图像，它的数学解析表达式由于书写起来过长已经难于在这里列出了。

当样本向量的输入维数大于二时，我们根本无法画出它的直观图形，只能靠类比来想象其非线性本质及强大的函数拟合能力。至于模型精确的数学解析表达式，采用如式（5.1）的紧凑而简洁的记法就是最好表达。

实际应用中样本向量的输入维数往往是几十、几百甚至更高的，支持向量的个数往往很多，SVM 决策函数的复杂性可以想见。

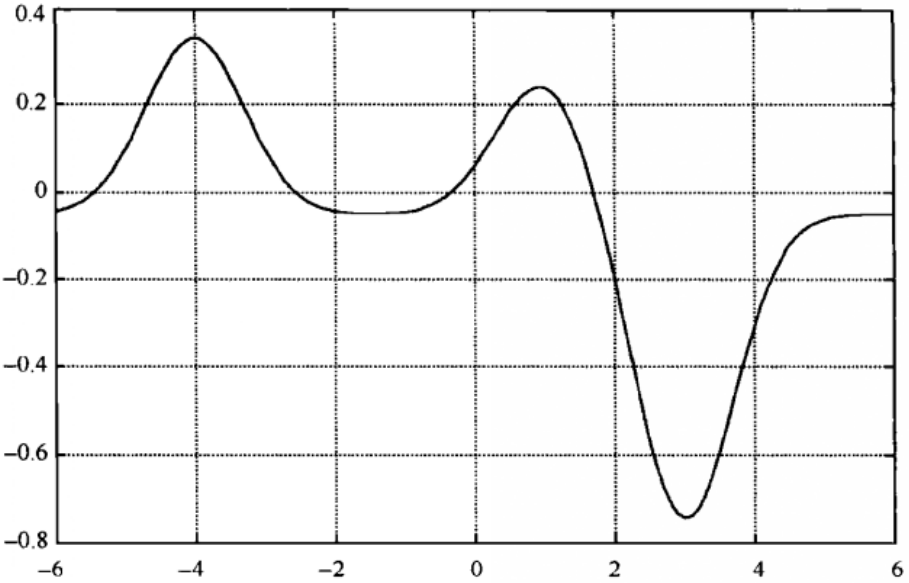


图 5.3　一维 SVM 决策函数图像

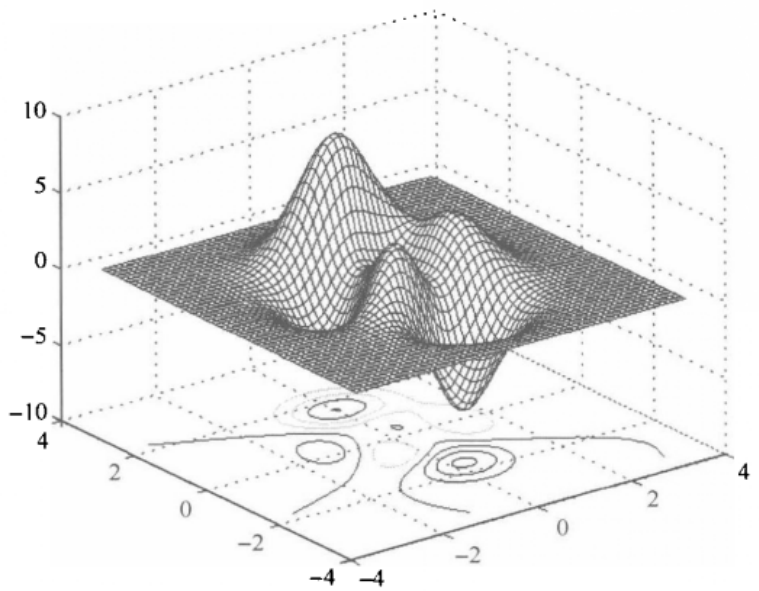


图 5.4　二维 SVM 决策函数图像

通过上述例子我们还可以看出：SVM 的最终决策函数实际上是多个非线性(核)函数的叠加结果。支持向量个数越多，叠加得到的最终决策函数越复杂。如果每一个样本都成了支持向量，得到的模型的拟合效果肯定好但过于复杂，导致推广泛化能力很差。

5.3　关于核方法

SVM 方法通常又称为基于核的方法。核方法研究是目前国际学术研究的新热

点之一。

几乎第 2 章讨论的全部内容，如最优划分超平面的概念和它的求解方法等，早在 1965 年就已经完整地提出了。由于它只能用来处理线性问题，所以一直未受重视。默默无闻了 27 年后，直到 1992 年人们才发现可以借助于 Mercer 定理(定理 3.1)引入核函数，用这一方法处理大量的非线性问题。这一扩展立刻受到人们的极大关注和重视，一系列新的理论和应用成果应运而生。因此人们称 SVM 方法为基于核的方法。

线性关系可以表示成点积的形式，而点积又确定了 Mercer 核[式(3.7)]，通过核函数的引入，就把原来的非线性关系转化为高维特征空间中的线性问题了。因此可以说：核方法具有变非线性为线性的能力，或者说它是能够用解线性问题的算法解决非线性问题。这一方法很容易推广到很多研究和应用领域，下面初步列举几项。

5.3.1 基于核的 SVM 判别分析

通常统计预报中的判别分析方法是求出一个判别方程，然后确定判别值。判别方程实际上是对应于一个超平面的，这类问题本质上是一个分类识别问题。

几乎可以不加修改地把 SVM 分类方法应用到判别分析中，而且得到的是复杂的非线性判别曲面，相对容易地解决非线性判别分析问题。

5.3.2 基于核的 SVM 时间序列分析

统计预报中各种时间序列分析方法实质上是一种线性回归问题，在不同的约束条件下寻求线性回归函数。

如果把第 3 章给出的 SVM 回归方法应用于时间序列分析，则得到的是各种非线性最优回归函数。

同样的思路可以应用到相关分析、MOS 等统计预报方法中。

5.3.3 基于 SVM 方法的数据修复

人们希望数据资料是连续和完整的，但是缺损个别数据是常有的事，特别是对于历史数据。所以，人们通常是采用线性插值的方法做数据修补，这是一种简单化的处理方法。

利用大量已有数据建立 SVM 回归函数，从而预测出缺损点的数据来，这不失为是一种数据修复的新方法，特别适合于处理复杂、多维和无法插值的数据修复。

5.3.4 基于 SVM 方法的数据挖掘

随着计算机应用水平的提高，需要处理的数据量迅速膨胀，信息压缩和数据挖掘成为必须面对的问题。这一问题的核心是如何从海量数据中提取有用的信息，从而抛弃冗余数据。SVM 方法在这方面可以有所作为。

首先，SVM 方法建立的模型只依赖于稀疏的样本(支持向量)，这为样本分析(抓住关键样本、剔除冗余样本)提供了重要信息。对于一些在训练学习中不管核函数、

核参数怎么变，都扮演支持向量角色的样本（通常称一个应用问题中的这类样本为本性支持向量），应该给予特殊的关注。特别要审核本来数量就很少的对应于小概率事件的样本点，看看他们是否在支持向量集合中。

其次，可以用 SVM 方法做因子分析。至少这可以从两方面进行：如果已建立的学习机效果理想，则可以利用已建立的学习机做删除因子实验；如果所建立的学习机效果很差——不管如何调整核和参数效果都很差，则反过来可以审核所选因子的合理性，增加新因子，删除不合理的因子。这也间接提醒我们：SVM 只是一种新的机器学习方法，它的应用效果好坏首先取决于样本数据因子的选取。

5.3.5　作为 SVM 的两层神经网络

我们选择核函数为

$$K(\boldsymbol{X},\boldsymbol{X}_i) = S(a(\boldsymbol{X}\cdot\boldsymbol{X}_i)+c),$$

则应用此核由训练样本构造的支持向量机决策函数为

$$f(\boldsymbol{X},\boldsymbol{\alpha}) = \operatorname{sgn}\Big(\sum_{i=1}^{n}\alpha_i S(a(\boldsymbol{X}\cdot\boldsymbol{X}_i)+c)+b\Big),$$

它实际上恰好是一个两层的神经网络。

其中 a 和 c 为参数，$S(u)$ 是如图 5.5 所示的 Sigmoid 函数，则应用此核由训练样本构造的 SVM 决策函数为

$$f(\boldsymbol{X},\boldsymbol{\alpha}) = \operatorname{sgn}\Big(\sum_{i=1}^{n}\alpha_i S(a(\boldsymbol{X}\cdot\boldsymbol{X}_i)+c)+b\Big),$$

它实际上恰好是一个两层的神经网络。图 5.6 给出了相应的图示。

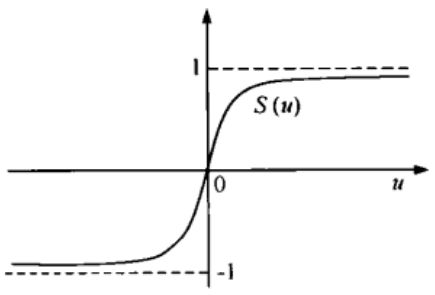


图 5.5　Sigmoid 函数

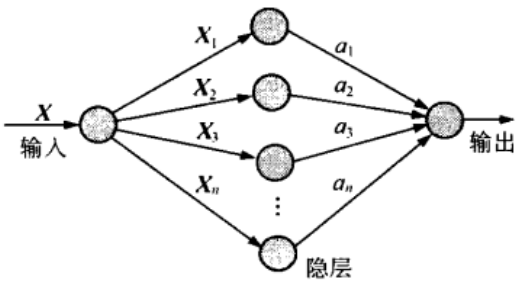


图 5.6　包含两层神经网的 SVM

在 SVM 两层神经网中，隐层单元的数目 n 恰好为支持向量个数；每个支持向量为第一层对应神经元的权值向量；每个支持向量的系数为第二层神经元的权值。普通两层神经网的隐层单元个数要事先指定，而这里却是自动优化计算确定的，这是 SVM 神经网络的优点之一。

5.3.6 用 SVM 方法求解算子方程

大量的线性问题可以归结为一种线性算子方程

$$\boldsymbol{A}f(x) = F(x),$$

其中 $F(x)$为已知函数，$f(x)$为待求函数，$\boldsymbol{A}$ 为一个线性算子。代数方程、微分方程、积分方程等都是线性算子方程的特例。SVM 方法给出了求解这类方程的新方案，特别是当方程的右边是形如式(4.1)给出的训练样本数据时。

5.4 学习机性能的评价

无论是分类问题还是回归问题，都存在一个对学习效果即学习机性能的评估问题。

人们习惯认为，只要建立的学习机对训练样本的累计误差最小，则学习机的效果最佳。这实际上是依据了一个所谓经验风险最小化(Empirical Risk Minimization，ERM)归纳原则，来最小化经验数据的风险泛函。

一般来说，为了能够得到最好的学习机，就需要在指定的函数类中选择对训练器响应最好的逼近函数。这就要考虑对于给定输入 $\boldsymbol{X}$ 得到的训练器响应 y，与学习机给出的响应 $f(\boldsymbol{X},\boldsymbol{\alpha})$之间的损失(误差)函数 $L(y, f(\boldsymbol{X},\boldsymbol{\alpha}))$，$\boldsymbol{\alpha}$ 为参数向量。如果给出的经验数据为

$$(\boldsymbol{X}_1, y_1), (\boldsymbol{X}_2, y_2), \cdots, (\boldsymbol{X}_l, y_l),$$

则这个损失函数可表示为 $L(y_i, f(\boldsymbol{X}_i,\boldsymbol{\alpha}))$。它的数学期望值为

$$R_{\text{emp}}(\boldsymbol{\alpha}) = \frac{1}{l}\sum_{i=1}^{l} L(y_i - f(\boldsymbol{X}_i,\boldsymbol{\alpha})) \tag{5.2}$$

这是一个基于参数 $\boldsymbol{\alpha}$ 的泛函，称为经验风险泛函。对于模式识别问题，$R_{\text{emp}}(\boldsymbol{\alpha})$就是错分率；对于回归分析问题，$R_{\text{emp}}(\boldsymbol{\alpha})$对应于均方差等不同的误差函数。

所谓 ERM 归纳原则就是选取 $f(\boldsymbol{X},\boldsymbol{\alpha})$，使 $R_{\text{emp}}(\boldsymbol{\alpha})$的值最小。然而对于学习问题来说，这一原则并不是最重要的。人们从训练样本数据求得最优函数的目的，主要不在于它是否可以拟合经验数据，而在于它对未来数据的预报效果，这就是学习机的推广能力。

单纯从函数拟合上看，对于任何给定的有限个经验样本数据，都不难找到一个(甚至是多个)函数，使其可以完全拟合这些数据，但很可能它们对于推广预报却毫无保证，有些甚至根本没有任何预报能力，图 5.7 给出一个这样的示例。图中样本点是二维数据，如图中的 15 个黑圆点。通过确定系数 k，找到了函数 $y=\sin(kx)$可以完全拟合这些样本点，但它对于实际问题毫无预报能力。实际上通过选择足够大的 k，函数 $y=\sin(kx)$可以充分拟合任何 y 值落在区间$[-1,1]$的有限样本数据。这类问

题在以往的学习方法选取过程中都会发生，在人工神经网研究中通常被称为所谓的“过学习”问题，即为了满足一定的拟合精度造成函数的复杂性加大，达到一定的限度后，$R_{\text{emp}}(\boldsymbol{\alpha})$越小泛化效果反而越差。

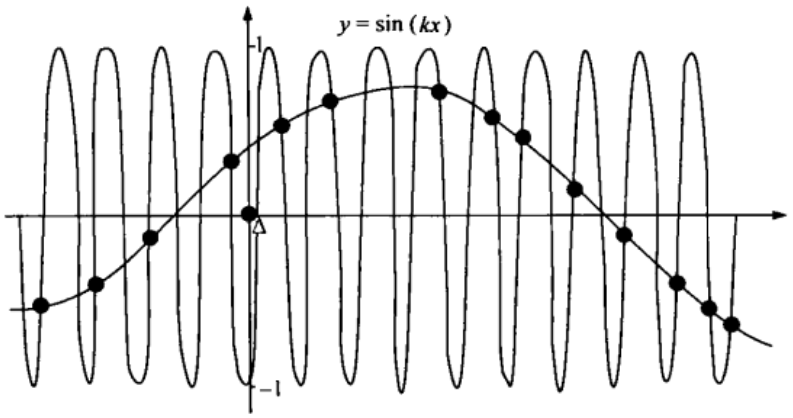


图 5.7　拟合好而无预报能力的示例

因此可以说——函数对样本拟合得完美未必推广（预报）能力就强，ERM 归纳原则受到了严重的挑战。SVM 方法出现以前的所有统计建模方法，都对于所得模型的推广能力无可奉告。如何评估一个学习模型的好坏？学习模型的推广能力究竟取决于什么？有没有办法估计所建立的模型的推广能力？这些都是机器学习理论的核心问题。

Vapnik(2004)提出的统计学习理论对于这些问题给出了十分出色的结果，严格证明了号称是机器学习理论“三个里程碑”的基本定理。

根据 Vapnik 的理论，决定学习机好坏的风险泛函，除了式(5.2)给出的经验风险项外，还应该有一项，它不受具体的回归函数影响，而是取决于备选函数集的某种整体性质。这里我们给出一个极端简化的表示式

$$R(\boldsymbol{\alpha}) \leqslant R_{\text{emp}}(\boldsymbol{\alpha}) + R_s(h) \tag{5.3}$$

式(5.3)左端为学习机的实际风险泛函，右端第一项为式(5.2)给出的经验风险泛函，它取决于函数集中的某一个特定的函数，右端第二项与选择具体函数的参数 $\boldsymbol{\alpha}$ 无关，它取决于称为 VC 维(Vapnik-Chervonenkis dimension)的参数 h。VC 维是一个新的数学概念，它反映备选函数集的某种整体容量特征。

这样，我们要极小化实际风险泛函，就必须对式(5.3)右边的两项同时最小化。对此，Vapnik 给出了一个称作结构风险最小化(Structural Risk Minimization，SRM)归纳原则。按照这一原则，决定一个学习机推广能力的主要因素，不是样本数据的好坏，也不是备选函数集性质的好坏(如参数的多少、光滑性等)，甚至也不是学习方法的异同。刻画备选函数集整体性质的 VC 维是决定学习机推广能力的关键因素，特

别是在训练数据是小样本的情况下，应用 VC 维可以给出学习训练所建模型的推广能力的估计。

第 6 章将对 VC 维、ERM 和 SRM 作较为详细地讨论。而第 7 章介绍的软件平台 CMSVM 2.0 就是依据 SRM 归纳原则，运行后自动给出所建模型的性能评估结果，给出对所建模型推广能力的估计。建议在熟悉 CMSVM 2.0 的使用后，格外注意这些事项。

5.5 标准 SVM 及其变种

前面几章介绍的 SVM 分类和回归方法通常称为基本 SVM 模型(或称标准 SVM)。作为国内外的研究热点之一，SVM 方法近年来在基本模型的基础上，又有很多新的进展，相继有人提出了许多新的 SVM 方法，如 ν-SVM、LS-SVM、1 模 SVM、单类 SVM 等。

这些多是理论上对基本模型在某一方面的纵横延拓，可能对某些问题带来一些好的效果，对应用效果并未带来实质性变化。从本节后面的简要介绍中可以看出，这些新模型在方法上也没有带来任何突破性的进展。

回顾一下基本 SVM 分类模型的最优化提法为

$$\min\left\{\frac{1}{2}\ \|\boldsymbol{W}\|^2+C\sum_{i=1}^{l}\xi_i\right\},$$

约束条件：$y_i((\boldsymbol{W}\cdot\boldsymbol{X}_i)+b)\geqslant 1-\xi_i \qquad (i=1,\cdots,l;\ \xi_i\geqslant 0)$。

将上述模型的某一方面加以合理扩展，就会产生一种新的 SVM 方法。

5.5.1 ν-SVM

基本 SVM 模型中的参数 C 的含义没有具体化，有人采用正则化的手段，试图给最优化中的参数以确定的物理含义，这对于参数优化能有帮助。将基本模型中的目标函数引入正则化参数 ν 和 ρ，原来最优化问题的目标函数变为

$$\min\left\{\frac{1}{2}\|\boldsymbol{W}\|^2-\nu\rho+\frac{1}{l}\sum_{i=1}^{l}\xi_i\right\},$$

其中 l 为样本数，ρ 为间隔值，ν 为人工可以调整的参数，用来控制允许错分样本的个数，$p/l\leqslant\nu\leqslant q/l$，$p$ 为错分样本数，q 为支持向量个数。这样模型中的参数就有了确定的含义，给优选模型和调整参数带来了方便。

5.5.2 LS-SVM

LS-SVM 又称最小二乘 SVM，主要是对原来的最优化问题的约束条件进行了修改，大于等于号变成了等号

$$\text{约束条件：}\ y_i((\boldsymbol{W}\cdot\boldsymbol{X}_i)+b)=1-\xi_i \qquad (i=1,\cdots,l),$$

如此修改导致 ξ_i 可能为负，所以又将目标函数中的 ξ_i 修改为 ξ_i^2。

这样的修改导致的一个结果是所有样本都将成为支持向量，不得不说在这一方面是一个退步。将其应用在某些实际的回归问题上据说得到了较好的精度。

5.5.3　1 模 SVM

在基本模型中，$\|\boldsymbol{W}\|$ 为一欧几里得距离。在一般的泛函空间中，距离是公理化定义的，可以给出不同的距离定义。即使是在同一实数集上也可以定义不同的距离。比如取绝对值 $|\boldsymbol{W}|$ 也是一种距离，一般的有 L^p 距离等。当 p 取 2 时，就是欧几里得距离；当 p 取 1 时，就称为 1 模，就是取绝对值距离。

将原来规划中的目标函数 $\|\boldsymbol{W}\|^2$ 改为 $|\boldsymbol{W}|$ 就变成了 1 模 SVM。这一变动可能对有些问题的解决表现了好的特性，但也带来了算法的复杂性。

除此之外，还有数目可观的 SVM 实现方法，但大多仍停留在论文研究阶段，未见有力的实际应用例证。

5.6　SVM 方法的弱点和“开问题”

任何方法和技术都有弱点，或说存在未解决的“开问题”，SVM 方法也不例外。SVM 方法又称为核方法或核机器，这是因为核的展开和计算是这一方法的关键。这一方法目前存在的最大问题是核函数的选定及有关核参数的优化，这是目前尚未解决的问题，也是制约 SVM 方法应用的主要因素。至今人们还不能回答——对于特定的数据集，是否存在最优核函数和最优核参数？如果有，如何选取？

此外，下面是 SVM 方法理论研究的关键问题：

(1) SRM 归纳原则是否有其他的实现途径？

(2) 所建模型泛化能力的界能否改进？

(3) Mercer 定理的限制能否突破？

对于气象业务应用来说，我们希望并期待这些问题早日得到解决，但不可能等待所有理论问题都解决了之后才开始应用该方法。结合近年在天气预报中的应用，我们对这一问题作了多方面的研究，除取得了一些优选参数的经验方法外，还开发出一个实用化的参数寻优软件平台 CMSVM，它是应用 SVM 方法的得力工具，应用它可以大大减轻核参数寻优的负担。多年来在气象台站的推广应用实践中证明，该软件平台的功能较为齐备，运行稳定可靠，操作使用方便。该软件平台不但能适应目前气象业务应用的需要，而且已经引起很多其他行业、科研院所和高等学校的重视。

5.7　本章小结

在掌握 SVM 分类与回归方法的基础上，本章对 SVM 方法的若干理论问题作了一些初步的分析，使人们能进一步加深对这一新方法的理解。主要包括：

(1) 比较了 SVM 方法与通常“先归纳，后演绎”推理方法间二者的不同，分析了

SVM 方法从样本到样本的直推建模的特点；

(2) 进一步分析了 SVM 模型的非线性本质；

(3) 简略介绍了作为核方法一种的 SVM 方法在其他方面的潜在应用，如基于 SVM 的判别分析、时间序列分析、数据修复、数据挖掘、神经网络、求解算子方程等；

(4) 推广能力大小是学习机性能好坏的主要标志，5.4 节初步涉及了 SRM 归纳原则和 VC 维的基本概念；

(5) 对于标准 SVM 的多种改进模型，如 ν-SVM，LS-SVM 和 1 模 SVM 等作了简要介绍；

(6) 简要分析了 SVM 方法的弱点和“开问题”，供理论研究工作者参考。

第6章　统计学习理论

前面几章已经通过分类和回归的建模问题初步介绍了 SVM 方法的基本思路。很显然，这是一种新颖的统计学习理论和方法，它的基本框架具有一般性。在理解 SVM 基本方法的基础上，本章从几个方面对统计学习的理论进行介绍。事实上，这里将要介绍的理论成果早在三十多年前就已经有人提出，SVM 方法不过是在这一理论基础上近年才研究发展出的应用方法。为了避免在一开始就涉及太多的数学问题而增加学习阅读的难度，我们在介绍上颠倒了它们的历史顺序。

相关数学理论可以参阅我们提供的文献资料。Vapnik V N (2004)给出了统计学习理论最完整的表述，在其中可以找到所有重要定理的证明。Vapnik V N (2000)是该书的简化版本，避开了全部数学证明而保留了主要结果①。

我们力图用通俗的语言介绍 VC 维和统计学习一致性的几个基本定理，目的是使 SVM 方法的应用者能够了解这一理论的轮廓，从而在实际应用中减少盲目性。

6.1　ERM 归纳原则

经典逻辑中的主要推理方法为演绎推理和归纳推理。归纳是描述从特殊到一般的过程，演绎则描述从一般到特殊的过程。从逻辑上看，机器学习主要是归纳过程。具体说来就是基于给定的经验数据(满足概率同分布)提供的有限信息，在允许的函数集中选出一个对经验数据的拟合最优者。所谓最优，即从传统的统计学习方法看来，就是要求学习所得的最优函数对于经验数据出错的风险最小，这导致一个基本原则——ERM 归纳原则，它是一个十分重要的原则。

设用于学习的经验数据

$$(\boldsymbol{X}_1, y_1), (\boldsymbol{X}_2, y_2), \cdots, (\boldsymbol{X}_l, y_l) \tag{6.1}$$

它们是根据同一联合分布 $F(\boldsymbol{X}, y) = F(\boldsymbol{X})F(y|\boldsymbol{X})$ 随机选取出的。此外还给定一个备选函数集 $\{f(\boldsymbol{X}, \boldsymbol{\alpha})|\boldsymbol{\alpha} \in \boldsymbol{\Lambda}\}$，$\boldsymbol{\alpha}$ 为参数。为了估计误差，需要事先指定一个度量经验数据输出 y 与学习机响应 $f(\boldsymbol{X}, \boldsymbol{\alpha})$ 之间差异的损失函数 $L(y, f(\boldsymbol{X}, \boldsymbol{\alpha}))$。

对于不同的问题可以指定不同的损失函数。比如，模式识别问题通常取损失函数为

$$L(y, f(\boldsymbol{X}, \boldsymbol{\alpha})) = \begin{cases} 0 & \text{当 } y = f(\boldsymbol{X}, \boldsymbol{\alpha}) \\ 1 & \text{当 } y \neq f(\boldsymbol{X}, \boldsymbol{\alpha}) \end{cases},$$

① 第 6 章有关理论内容主要参考了这两本著作，VC 维的相关内容参见 Vidyasagar M(1997)。

由于这类函数只取 0,1 二值,分别对应正确和错误的分类,所以在统计学习中它们被称为指示函数(不要小看这类简单函数,很多有关实函数的问题都可以通过先解决指示函数,进一步把实函数表示成指示函数从而得到解决。统计学习理论的很多证明都是如此处理)。在回归分析中可以取 4.2 节提到的各种误差函数作为损失函数,如平方函数

$$L(y,\ f(\boldsymbol{X},\boldsymbol{\alpha})) = (y - f(\boldsymbol{X},\boldsymbol{\alpha}))^2。$$

由于经验数据的产生带有随机性,所以估计误差时必须考虑损失函数的数学期望值

$$R(\boldsymbol{\alpha}) = \int L(y, f(\boldsymbol{X},\boldsymbol{\alpha}))\mathrm{d}f(\boldsymbol{X},y) \tag{6.2}$$

$R(\boldsymbol{\alpha})$就是反映学习机关于经验数据的响应好坏的指标,通常称之为风险泛函。传统学习的目的就是要在分布函数 $F(\boldsymbol{X},y)$未知的情况下,只根据经验数据[如式(6.1)]来最小化风险泛函。具体说来就是在备选的函数类$\{f(\boldsymbol{X},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$中寻找一个最优函数 $f(\boldsymbol{X},\boldsymbol{\alpha}_0)$,它能够使得 $R(\boldsymbol{\alpha})$取得最小。

为了使讨论能包括密度函数估计等各种具体的学习问题,我们用$\{\boldsymbol{z}_i\mid i=1,\cdots,l\}$表示式(6.1)中的数据$\{(\boldsymbol{X}_i,y_i)\}$,$\boldsymbol{z}$ 为$(n+1)$维向量

$$\boldsymbol{z}_1,\ \boldsymbol{z}_2,\cdots,\ \boldsymbol{z}_l \tag{6.3}$$

用 $Q(\boldsymbol{z},\boldsymbol{\alpha})$表示损失函数,则(6.2)式的风险泛函可以简化表示为

$$R(\boldsymbol{\alpha}) = \int Q(\boldsymbol{z},\boldsymbol{\alpha})\mathrm{d}F(\boldsymbol{z}) \tag{6.4}$$

由此,学习问题就转变成求能使得 $R(\boldsymbol{\alpha})$最小的 $Q(\boldsymbol{z},\boldsymbol{\alpha}_0)$。

由于分布函数 $F(\boldsymbol{z})$是未知的,所以上式的风险泛函实际上是无法实际计算的。但我们根据已知的经验数据$\{\boldsymbol{z}_i\}$,可以建立如下的经验数据的误差期望值(平均值)

$$R_{\mathrm{emp}}(\boldsymbol{\alpha}) = \frac{1}{l}\sum_{i=1}^{l} Q(\boldsymbol{z}_i,\boldsymbol{\alpha}) \tag{6.5}$$

式(6.5)被称为经验风险泛函,它是由训练数据集直接算出的。寻求 $Q(\boldsymbol{z},\boldsymbol{\alpha}_i)$使得式(6.5)最小化的归纳原则称为经验风险最小化(Empirical Risk Minimization,ERM)归纳原则,该原则在第 5 章曾提及。经验风险泛函 R_{emp}的值一般依赖于经验样本集和不同的函数 $Q(\boldsymbol{z},\boldsymbol{\alpha})$。

$R_{\mathrm{emp}}(\boldsymbol{\alpha})$是对 $R(\boldsymbol{\alpha})$的近似和逼近。通常认为小的 $R_{\mathrm{emp}}(\boldsymbol{\alpha})$一定对应小的 $R(\boldsymbol{\alpha})$,使 $R_{\mathrm{emp}}(\boldsymbol{\alpha})$最小的学习机就是性能最优。

按照这种归纳原则,提高学习机的性能有两个努力方向:一是扩大备选函数集,

更大范围的最优总不会比小范围的最优差；二是使学习过程在较小的训练样本集上可以实现，因为实际问题都是小样本的。

ERM 归纳原则在传统的学习理论中扮演着决定性的角色，工程计算中常用的最小二乘法和概率中的最大似然估计法等都是基于这一原则建立的。然而，这里存在如下几个问题：

(1) 寻宝最好事先知道宝是存在的，否则便是盲目行动。ERM 归纳原则是要在指定函数集中寻求最优响应函数，不免也带有盲目性——所寻求的最优函数在指定函数集中是否存在？若存在，这种归纳寻优的算法是否可以保证逐渐靠近它？若逐渐靠近，靠近的速度是否很快？速度是否可控？这一系列问题 ERM 无法回答。

(2) 正如归纳的目的是要演绎一样，学习的目的是要推广。ERM 归纳原则关心的是对经验数据的拟合误差，对未来的推广能力无可奉告，这是这一归纳原则的致命伤。可以列举很多例子说明，对经验数据拟合得好并不一定推广效果好，依据这一归纳原则得出的最优有时反而不优。

Vapnik 等提出的新的统计学习理论对上述问题给出了出色的回答。对于问题(1)，他们定义了学习过程的一致性，进而严格证明了关于学习过程一致性的一系列基本定理，包括三个充分必要条件，这三个条件号称是解决学习理论问题的里程碑。对于问题(2)，Vapnik 和 Chervonenkis 引入了刻画函数集容量的 VC 维概念，提出了一个新的结构风险最小化（Structural Risk Minimization，SRM）归纳原则，5.4 节曾对其进行粗略地描述。其使得人们可以对学习机的推广能力进行估计，并给出了学习机误差的界。

新的统计学习理论证明了，决定学习机推广能力的，不单是由经验数据算得的 $R_{\text{emp}}(\boldsymbol{\alpha})$，还有与经验数据无关的备选函数集的“容量”。Vapnik 等得出的这些成果奠定了一座新的统计学习理论的基石。

6.2　统计学习的基本定理

统计学习理论所要解决的核心问题是，对一个依据 ERM 归纳原则建立的学习过程，什么条件下能够取得小的实际风险，什么情况下不能。学习过程的这一特性称为“学习过程的一致性”，它需要通过样本数 l 无限变大的渐进过程来检验。在数学上可以通过极限给出如下定义：

定义 6.1　给定数据如式(6.3)，$R(\boldsymbol{\alpha})$ 和 $R_{\text{emp}}(\boldsymbol{\alpha})$ 分别是式(6.4)和(6.5)给出的期望风险泛函和经验风险泛函，$Q(\boldsymbol{z},\boldsymbol{\alpha}_l)$ 是使 $R_{\text{emp}}(\boldsymbol{\alpha})$ 最小化的函数。如果下面两个序列依概率收敛于同一个极限，即

$$\lim_{l\to\infty(p.)} R(\boldsymbol{\alpha}_l) = \inf_{\boldsymbol{\alpha}\in\Lambda} R(\boldsymbol{\alpha}) \tag{6.6}$$

$$\lim_{l\to\infty(p.)} R_{\text{emp}}(\boldsymbol{\alpha}_l) = \inf_{\boldsymbol{\alpha}\in\Lambda} R(\boldsymbol{\alpha}) \tag{6.7}$$

则说 ERM 归纳原则对于函数集$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$和概率分布函数 $F(\boldsymbol{z})$是一致的。公共的极限值为最小可能的风险值。式(6.6)保证了所达到的风险收敛于最好的可能值,而式(6.7)保证了可以用经验风险的值逼近和估计最小可能的风险值。图 6.1 给出了直观描述。

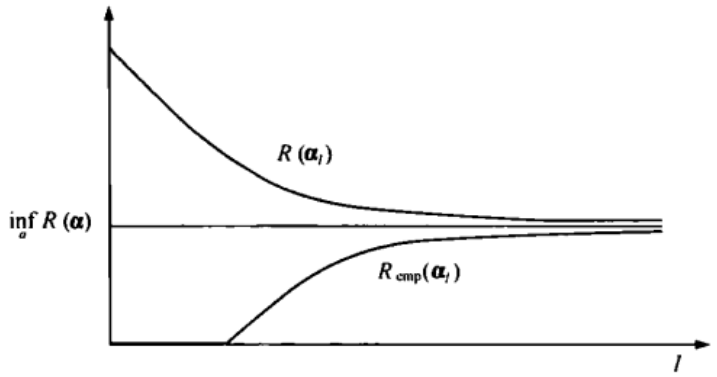


图 6.1 ERM 学习一致性图示

定义 6.1 给出一个统一的判断标准。基于这一定义,Vipnik 证明了学习理论的如下关键定理:

定理 6.1 设函数集$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$满足条件

$$A\leqslant\int Q(\boldsymbol{z},\boldsymbol{\alpha})\mathrm{d}F(\boldsymbol{z})\leqslant B,$$

则 ERM 归纳原则一致性的充分必要条件是:经验风险 $R_{\text{emp}}(\boldsymbol{\alpha})$在函数集$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\Lambda\}$上,如下一致收敛于实际风险 $R(\boldsymbol{\alpha})$

$$\lim_{l\to\infty}P\{\sup_{\boldsymbol{\alpha}\in\Lambda}(R(\boldsymbol{\alpha})-R_{\text{emp}}(\boldsymbol{\alpha}))>\varepsilon\}=0\qquad(\forall\varepsilon>0)\tag{6.8}$$

这种一致收敛称为一致单边收敛(有别于差式加上绝对值号的一致双边收敛)。我们对于这一关键定理略加说明:

(1) 这里给出的是一个很强的充分必要条件,定义 6.1 界定的 ERM 归纳原则的一致性等价于一致单边收敛;

(2) 这里的收敛是依概率的收敛,所关心的概率值是关于"最坏"的函数(使风险差值取最大(sup)的函数);

(3) 注意这里式(6.8)要求对于任意小正数 ε 均成立;

(4) 这一定理不是可构造的,但它的否定形式可以用来构造不一致的反例;

(5) 需要建立使式(6.8)成立的充分必要条件,进而考虑什么条件下它收敛的速度是快的,什么条件下它的收敛特性与概率测度的定义无关。这导致下面的三个重要定理。

为了避免涉及更多的数学概念,本节只考虑模式识别的情况,即设函数集

$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$为指示函数集。

定义 6.2　(VC 熵)给定指示函数集$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$和数据[如式(6.3)],定义量$N^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l)$为应用指示函数集中的函数能够把样本数据式(6.3)分成不同的子集的个数;称值 $H^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l)=\ln(N^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l))$为给定指示函数集关于给定数据的随机熵;称随机熵在联合分布函数 $F(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l)$上的期望

$$H^{\Lambda}(l)=E(\ln(N^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l)))=E\ln N^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l),$$

为指示函数集在数量为 l 的样本集上的 VC 熵。

"熵"原本是统计物理学中的一个概念,用来刻画一个物理状态的分子热运动的平均紊乱程度,熵越大,紊乱程度越高。后来这个概念在很多方面被借用,如信息论中刻画平均信息量的"信息熵",模糊逻辑中刻画模糊系统或模糊集合的模糊程度的"模糊熵"等。一般称一种量的对数函数为其熵函数。这里的随机熵也是借用这个概念,来刻画一个指示函数集对随机产生的经验样本集进行分类的多样性。

从定义可以看出,VC 熵依赖于指示函数集、经验样本数 l 和概率测度(计算概率期望 E),它反映了给定指示函数集在数目为 l 的样本集上期望分类的多样性。

由于数目为 l 的样本集最多有 2^l 个不同的子集,所以成立关系式 $H^{\Lambda}(l)\leqslant l\ln 2$。

定义 6.3　如定义 6.2 所述,设 $N^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l)$为应用指示函数集中的函数能够把样本数据式(6.3)分成不同的子集的个数;称值

$$H^{\Lambda}_{ann}(l)=\ln(E(N^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l)))=\ln EN^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l)$$

为指示函数集在数量为 l 的样本集上的退火的 VC 熵。

所谓"退火的(Annealed)",是指交换了 VC 熵中取对数和取期望值的顺序,先取期望起了平滑作用。退火的 VC 熵同样依赖于指示函数集、经验样本数 l 和概率测度,它也反映给定指示函数集在数目为 l 的样本集上期望分类的多样性。

定义 6.4　如定义 6.2 所述,设 $N^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l)$为应用指示函数集中的函数能够把样本数据式(6.3)分成不同的子集的个数;称值

$$G^{\Lambda}(l)=\ln(\sup_{x_1,\cdots,x_l}(N^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l)))=\ln\sup_{x_1,\cdots,x_l}N^{\Lambda}(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_l),$$

为指示函数集在数量为 l 的样本集上的生长函数。

生长(Growth)函数也反映给定指示函数集在数目为 l 的样本集上分类的多样性,它只依赖于指示函数集和经验样本数 l,不依赖于概率测度。

从上述三个定义不难看出,三者有如下关系:

$$H^{\Lambda}(l)\leqslant H^{\Lambda}_{\text{ann}}(l)\leqslant G^{\Lambda}(l)\tag{6.9}$$

利用这里的三个定义,可以得到下面三个重要定理:

定理 6.2　经验风险 $R_{\text{emp}}(\boldsymbol{\alpha})$ 一致收敛于实际风险 $R(\boldsymbol{\alpha})$的充分必要条件是

$$\lim_{l\to\infty}\frac{H^{\Lambda}(l)}{l}=0 \tag{6.10}$$

这一定理号称是统计学习理论中的第一个里程碑，其中的 $H^{\Lambda}(l)$ 为 VC 熵。ERM 归纳原则要成立，即定理 6.1 中的式(6.8)要成立，首先必须满足这一定理所包含的条件。但它对于收敛的速度没有任何保证。我们可以找到收敛但收敛的渐进速度非常慢的特例。对于应用来说，这样的收敛也是没有意义的。

定理 6.3 经验风险 $R_{\text{emp}}(\boldsymbol{\alpha})$ 一致收敛于实际风险 $R(\boldsymbol{\alpha})$，且收敛速度快的充分条件是

$$\lim_{l\to\infty}\frac{H^{\Lambda}_{ann}(l)}{l}=0 \tag{6.11}$$

这一定理号称是统计学习理论中的第二个里程碑，其中的 $H^{\Lambda}_{ann}(l)$ 为退火的 VC 熵。它保证了 ERM 归纳原则成立且有快的收敛速度。注意，这里只是充分条件，必要条件至今尚未找到。

VC 熵和退火的 VC 熵都依赖于给定的概率测度，这从它们的定义就可以看出。能否找到 ERM 是一致的、收敛快而又不受此限制的条件呢？

定理 6.4 经验风险 $R_{\text{emp}}(\boldsymbol{\alpha})$ 一致收敛于实际风险 $R(\boldsymbol{\alpha})$ 且不依赖于概率测度的充分必要条件是

$$\lim_{l\to\infty}\frac{G^{\Lambda}(l)}{l}=0 \tag{6.12}$$

这一定理号称是学习理论中的第三个里程碑，其中的 $G^{\Lambda}(l)$ 为生长函数。它保证了 ERM 归纳原则成立且与概率测度的定义无关。

由式(6.9)所表述的关系可知：定理 6.4 的约束性最强，若式(6.12)成立，则式(6.10)和(6.11)自然成立。

上述三个定理解决了机器学习的最基本的理论问题，具有十分重要的理论意义。Vapnik V N(2004)中有这些定理的严格证明。但这些定理都不是构造性的，或者说无法拿来计算、建模和应用。对于实际问题，无法具体计算这些极限。

Vapnik 通过定义函数集的 VC 维，进而给出了一种可构造性计算的判定方法，并利用此方法可估计学习机推广能力的界。

6.3 函数集的 VC 维

VC 维是描述一个函数集的“容量”特性的抽象数学概念。我们先讨论最简单的函数集——指示函数集的 VC 维，进而把它推广到任意实函数集。

6.3.1 指示函数集的 VC 维

指示函数又称集合的特征函数，给定一个指示函数就对应一个由函数值为 1 的

点组成的集合。图 6.2 给出指示函数 $f_A(z)$与集合 A 的这种对应关系的直观图示。

因此，给定一个指示函数的集合 F，也就等价于给定了一个集合类 S。定义指示函数集合 F 的 VC 维也就等价于定义与它对应的集合类 S 的 VC 维。下面先给出其严格定义，再加以解释：

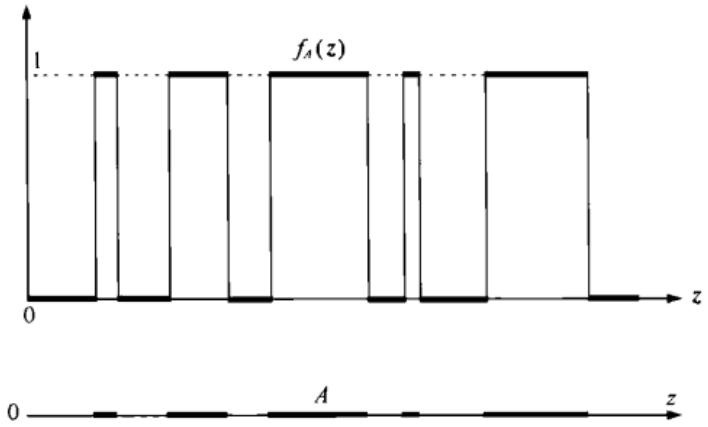


图 6.2　指示函数 $f_A(\boldsymbol{z})$与对应的集合 A

定义 6.5　设 $\boldsymbol{Z}=\{z_1,z_2,\cdots,z_l\}$是一个给定集合，$\boldsymbol{S}$ 是一个集合类。如果对于 $\boldsymbol{Z}$ 的任何子集 $z\subseteq\boldsymbol{Z}$，都存在 $\boldsymbol{s}\in\boldsymbol{S}$，使得 $\boldsymbol{Z}\cap\boldsymbol{s}=\boldsymbol{z}$ 成立，则称集合 $\boldsymbol{Z}$ 被集合类 $\boldsymbol{S}$ 所“打散”。

集合类 $\boldsymbol{S}$ 所能“打散(Shattered)”的所有集合的最大基数称为集合类 $\boldsymbol{S}$ 的 VC 维。

集合类 $\boldsymbol{S}$ 的 VC 维也就是与之对应的指示函数集 F 的 VC 维。一个函数集 F 的 VC 维为 h，记为 VC-dim$(F)=h$。

VC 维是一个较为抽象的概念，理解它的关键是理解“打散”的含义。可以把函数集 F 的 VC 维理解为刻画 F 的“丰富程度(Richness)”的度量。集合 $\boldsymbol{Z}$ 可以被函数集 F 所“打散”，是指 F 的成员足够“丰富”，可以区别出 $\boldsymbol{Z}$ 的所有可能的子集。因此，也许用“完全区分”或“完全区别”代替“打散”会更好理解些。但人们对“打散”已经先入为主了。

如果 $\boldsymbol{Z}$ 是含有 l 个元素的有限集合，则集合 $\boldsymbol{Z}$ 含有 2^l 个子集。因此函数集 F 要想把 $\boldsymbol{Z}$“打散”，则它至少要含有 2^l 个不同的函数。反之，若 F 为有限集，则能够被 F “打散”的集合至多含有$[\log_2|F|]$个元素。这告诉我们——含有 l 个元素的有限函数集 F 的 VC 维是有限的，且有

$$\text{VC-dim}(F)\leqslant[\log_2|F|]。$$

无穷函数集的 VC 维可以是有限，也可以是无穷的。

定义 6.5 告诉我们：若已知 VC-dim$(F)=h$，则任何含有$(h+1)$个或更多不同元素的集合都不能被 F“打散”；但 F 必可以“打散”某一个含有 h 个元素的集合。

请注意上面的逻辑量词。定义 6.5 给出了判断 VC 维的直接方法：

(1) 若确认 VC-dim(F)$\leqslant h$，则必须说明任何含有($h+1$)个或更多不同元素的集合都不能被 F“打散”；

(2) 若确认 VC-dim(F)$\geqslant h$，则只需找出一个含有 h 个元素的集合，说明它能被 F“打散”。

如果上述两点同时都得到确认，则证明了 VC-dim(F)等于 h。

6.3.2 任意实函数集的 VC 维

任意实函数集的 VC 维的概念可以借助指示函数集的 VC 维的定义获得。实现这一扩充需要引入下面的概念：

定义 6.6 设实函数集$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$满足条件

$$A\leqslant\int Q(\boldsymbol{z},\boldsymbol{\alpha})\mathrm{d}F(\boldsymbol{z})\leqslant B,$$

式中，记区间(A,B)为 $\mathscr{B}$，其中 A 可以是$-\infty$，而 B 可以是$+\infty$。

称指示函数集$\{I(\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\beta})\}$是实函数集$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$的一个指示器集合，且

$$I(\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\beta})=\theta(Q(\boldsymbol{z},\boldsymbol{\alpha})-\boldsymbol{\beta})\qquad(\boldsymbol{\alpha}\in\boldsymbol{\Lambda},\ \boldsymbol{\beta}\in\mathscr{B}),$$

其中 $\theta(z)$为阶跃函数，当 $z\geqslant 0$ 时取值为 1，其余情况下取值为 0，即

$$\theta(z)=\begin{cases}0 & \text{当 } z<0\\ 1 & \text{当 } z\geqslant 0\end{cases}。$$

如定义 6.6，一个实函数 $Q(\boldsymbol{z},\boldsymbol{\alpha})$对应一系列指示函数 $\theta(Q(\boldsymbol{z},\boldsymbol{\alpha})-\boldsymbol{\beta})$(如图 6.3 所示)。

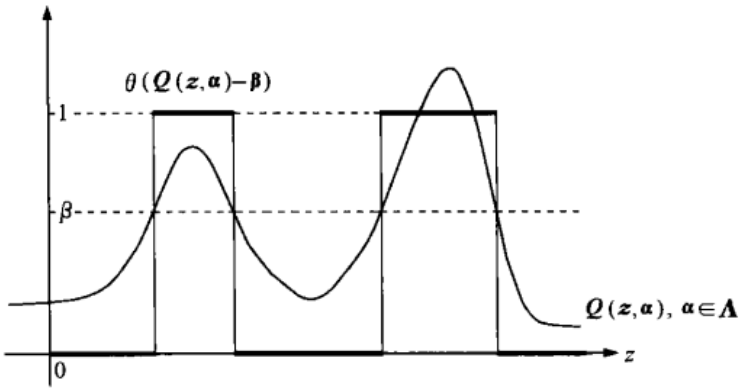


图 6.3 函数 $Q(\boldsymbol{z},\boldsymbol{\alpha})$与其一个指示函数

随着 $\boldsymbol{\beta}$ 取遍 $\mathscr{B}$ 中所有不同的值，就得到了与这个实函数对应的一系列指示函数；随着 $\boldsymbol{\alpha}$ 取遍 $\boldsymbol{\Lambda}$ 中所有不同的值，就得到了与这个实函数集合$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$相对应

的指示器集合(指示函数集合)。

定义 6.6 表明:任意一个实函数集合 F,均可以表现为一个指示器集合即指示函数集合 I。这样我们可以通过相应的指示函数集合的 VC 维来定义任意一个实函数集合的 VC 维。

定义 6.7　设实函数集 $F=\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$ 满足条件

$$A\leqslant\int Q(\boldsymbol{z},\boldsymbol{\alpha})\mathrm{d}F(\boldsymbol{z})\leqslant B,$$

式中,称指示函数集合 I 的 VC 维为此实函数集的 VC 维,记为 VC-dim(F)。其中 I 为定义 6.6 中所指的指示器集合。

可以验证,当实函数集 F 转化为指示函数集合时,即 $\mathscr{B}=\{0,1\}$,则定义 6.7 就与定义 6.5 等同。

这里我们得到了通过借助函数集的指示器集合 I 定义任意实函数集的 VC 熵、退火 VC 熵和生长函数的一种方法,我们同样把一个函数集的指示器的 VC 熵、退火 VC 熵和生长函数称为这个实函数集的 VC 熵、退火 VC 熵和生长函数,分别记为 $H^{\Lambda,\mathscr{B}}(l)$,$H_{\mathrm{ann}}^{\Lambda,\mathscr{B}}$ 和 $G^{\Lambda,\mathscr{B}}(l)$,这里 $\mathscr{B}$ 是函数集的函数取值范围。

同样,6.2 节中的三个重要定理(定理 6.2,6.3 和 6.4)对于任意实函数集也成立。

6.3.3　函数集 VC 维的例

在本小节我们通过分析一些特殊函数集 VC 维的例子,加深对 VC 维的理解。因为下面要判断给出的函数集例子的 VC 维,这里还要强调两点:

(1) 若一个函数集的 VC 维是 h,根据定义 6.5,只要求至少有一个含有 h 个元素的点集可以被它“打散”;并不要求任意一个含有 h 个元素的点集均可以被它“打散”。

(2) 若称一个函数集的 VC 维不低于 h,意味着至少可以找到一个含有 h 个元素的点集,可以被它“打散”;若称一个函数集的 VC 维不高于 h,意味着任意含有 $(h+1)$ 个元素的子集,均不能被它“打散”。验证一个函数集的 VC 维是 h,必须同时验证上述两个方面。读者可以对下面的例子自己加以验证。

例 6.1　设 A 为实直线 $\boldsymbol{R}$ 上所有半开区间 $(-\infty,t]$ 的集合,可以验证:VC-dim$(A)=1$。

例 6.2　设 A 为实直线 $\boldsymbol{R}$ 上所有闭区间 $[a,b]$ 的集合,可以验证:VC-dim$(A)=2$。

(注:若 A 为 Cantor 集的余集构成的集合,则 VC-dim$(A)=+\infty$)

例 6.3　设 A 为实平面 $\boldsymbol{R}^2$ 上所有半平面的集合,即

$$A=\{(x_1,x_2)\mid w_1x_1+w_2x_2-b\geqslant 0\},$$

可以验证:VC-dim$(A)=3$。

从例 6.3 可以看出的：实平面 $\boldsymbol{R}^2$ 上所有半平面的集合与 $\boldsymbol{R}^2$ 上所有线性函数的集合是对等的。

例 6.4 设 F 为实平面 $\boldsymbol{R}^2$ 上所有线性函数的集合，即

$$F = \{l \mid w_1 x_1 + w_2 x_2 - b = 0\}$$

可以验证：VC-dim$(F)=3$。

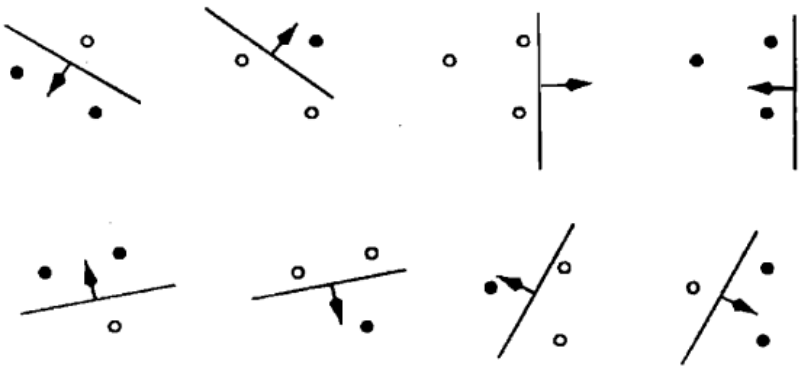


图 6.4 $\boldsymbol{R}^2$ 上所有直线“打散”了含有 3 个元素的点集

下面通过图示验证例 6.3 和 6.4 的结论。

根据前面定义 VC-dim$(F)=3$，说明平面上的所有直线所能“打散”的“最大”点集至多含有 3 个元素(点)。图 6.4 给出“打散”这个点集的直观图，它的 8 个子集均可由一条直线划分出，从中我们找到了一个含有 3 个元素的集合，它可以被函数集 F “打散”，这样有 VC-dim$(F)\geqslant 3$。另一方面，图 6.5 说明任何含有 4 个元素的点集不能被函数集 F “打散”，无法用一条直线划分出子集$\{z_2, z_4\}$，这说明 VC-dim$(F)\leqslant 3$。综合这两方面可以确认，实平面 $\boldsymbol{R}^2$ 上所有线性函数的集合的 VC 维是 3。

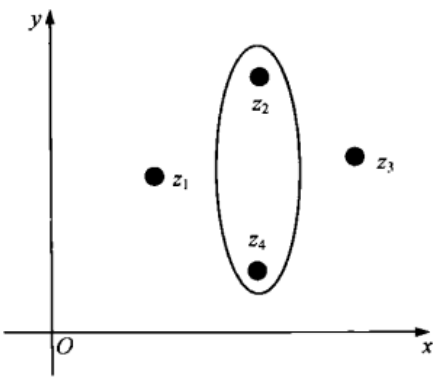


图 6.5 $\boldsymbol{R}^2$ 上所有直线不能“打散”的含 4 个元素的点集

需要注意的是，同在一条直线上的 3 个点就不能被这个函数集“打散”，但这并不妨碍上述结论的正确性。

与上面两例相似，下面分析实空间 $\boldsymbol{R}^3$ 中的情况。

例 6.5 设 A 为实空间 $\boldsymbol{R}^3$ 中所有半空间的集合(或 F 为实空间 $\boldsymbol{R}^3$ 中所有线性

函数的集合)，即

$$A = \{(x_1, x_2, x_3) \mid w_1 x_1 + w_2 x_2 + w_3 x_3 - b \geqslant 0\},$$

可以验证：VC-dim$(A) = 4$(或 VC-dim$(F)=4$)。

类似推广到高维空间，可以证明如下结果：

例 6.6　设 F 为实空间 $\boldsymbol{R}^n$ 中所有线性超平面(或半超平面)的集合(或 F 为实空间 $\boldsymbol{R}^n$ 中所有线性函数的集合)，即

$$F = \{\boldsymbol{X} \in \boldsymbol{R}^n \mid w_1 x_1 + w_2 x_2 + \cdots + w_n x_n - b = 0\},$$

可以验证：VC-dim$(F) = n+1$。

由此得出，2 维空间的所有线性函数的 VC 维为 3，3 维空间的所有线性函数的 VC 维为 4，n 维空间的所有线性函数的 VC 维为$(n+1)$。似乎 VC 维与函数的自由参数的个数有关，但这纯属巧合。图 6.6 所示事例中的函数只含有 1 个参数，但它的 VC 维是无穷大。

例 6.7　设 F 为实空间 $\boldsymbol{R}^2$ 中的函数集

$$F = \{f \mid y = \sin kx,\ k \in \boldsymbol{R}\},$$

则 VC-dim$(F)=+\infty$。

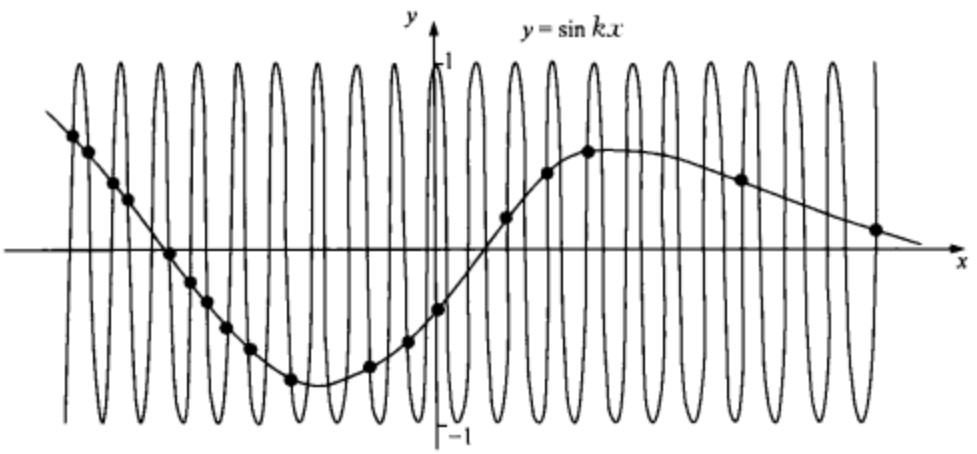


图 6.6　只含 1 个自由参数但 VC 维为无穷的函数集

按照定义 6.7 可以验证函数集 F 的 VC 维为无穷。因为对于任意大的整数 h，它都可以“打散”含有 h 个元素的点集，即它的 VC 维可以大过任意大的整数。

例 6.8　设 A 为实平面 $\boldsymbol{R}^2$ 上所有边平行于坐标轴的矩形，则

$$\text{VC-dim}(A) = 4$$

例 6.9　设 A 为空间 $\boldsymbol{R}^3$ 上所有面平行于坐标面的平行六面体，则

$$\text{VC-dim}(A) = 6。$$

例 6.10 设 A 为 $\boldsymbol{R}^n$ 上所有面平行于坐标面的超平行六面体，则

$$\text{VC-dim}(A) = 2n。$$

例 6.11 设 A 为[0,1]×[0,1]中的所有凸多边形的集合，则

$$\text{VC-dim}(A) = +\infty。$$

为了说明这一结论，只需要说明对于任意大的整数 n，都可以找到一个集合 B，使它所含的元素个数为 n，但它能够被某一个凸多边形完全区分出来，即 B 可以被 A 所“打散”。如图 6.7，设 B 的所有点在[0,1]×[0,1]内的一个圆周上，对于 B 的任意一个子集，则以这些点为顶点的凸多边形就可以把这些点区分出来。

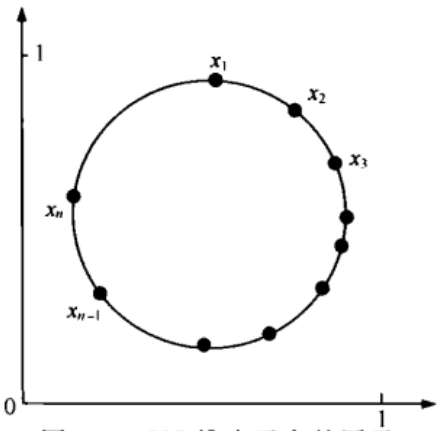


图 6.7 VC 维为无穷的图示

VC 维是一个函数集所具有的特性，它直接影响一个学习机的推广性能。应用它可以构造可计算的推广能力的界。

6.4 VC 维与生长函数的关系

应用生长函数我们得到了机器学习的第三个里程碑——ERM 一致收敛、收敛快且与概率测度的定义无关的充分必要条件(详见定理 6.4)。遗憾的是生长函数是不便于构造具体算法的。利用 VC 维与生长函数的关系我们可以补充上这一缺憾。

VC 维与生长函数之间有如下几个重要的定理。

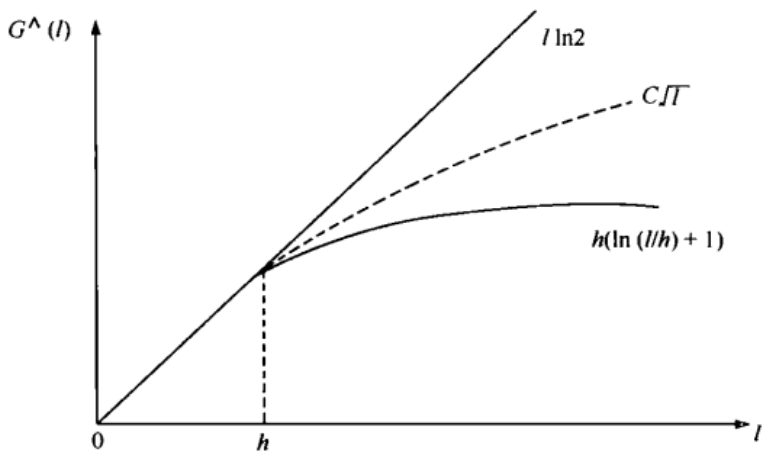


图 6.8 生长函数与 VC 维的关系

定理 6.5 任何生长函数，它或者满足等式

$$G^{\Lambda}(l) = l\ln 2,$$

或者受限于不等式：

$$G^{\Lambda}(l) \leqslant h\left(\ln\frac{l}{h}+1\right), \tag{6.13}$$

其中 h 是一个整数，使当 $l \leqslant h$ 时，有 $G^{\Lambda}(l)=l\ln 2$，而 $G^{\Lambda}(h+1)<(h+1)\ln 2$。

这一定理说明，生长函数要么是线性的函数，要么以一个对数函数为上界，如图 6.8 所示。

生长函数不可能是其他函数，比如不可能是图中虚线所示的幂函数等。h 是使等号成立的最大整数。由此我们得到了 VC 维的又一个定义。

定义 6.8　如果指示函数集$\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$的生长函数是线性的，则称这个函数集的 VC 维是无穷大；如果这个指示函数集的生长函数以参数为 h 的对数为界，则称这个函数集的 VC 维是有限的且等于 h。

可以证明这个定义与定义 6.5 是等价的。联系式(6.9)，我们得到了重要的不等式

$$\frac{H^{\Lambda}(l)}{l} \leqslant \frac{H^{\Lambda}_{\text{ann}}(l)}{l} \leqslant \frac{G^{\Lambda}(l)}{l} \leqslant \frac{h\left(\ln\frac{l}{h}+1\right)}{l} \qquad (l>h) \tag{6.14}$$

由此可以推出以下两个重要结论：

(1) 对于任何有限的 VC 维 h，容易证明(对函数 $y=\frac{\ln x}{x}$应用洛必达(L'Hospital)法则求极限)

$$\lim_{l\to\infty}\frac{h\left(\ln\frac{l}{h}+1\right)}{l}=0。$$

这意味着不等式(6.14)左边 3 项同样以 0 为极限，根据定理 6.2，6.3 和 6.4，经验风险 $R_{\text{emp}}(\boldsymbol{\alpha})$ 一致、快速、与概率测度的定义无关地收敛于实际风险 $R(\boldsymbol{\alpha})$。这样我们就找到了一个十分简单的判断归纳一致性的充分条件——函数集的 VC 维 h 为有限数。

(2) 若函数集的 VC 维为无穷，则 $G^{\Lambda}(l)=l\ln 2$ 成立，此时无法保证定理 6.2，6.3 和 6.4 的条件成立，ERM 归纳原则可能不成立(可以举出应用 ERM 归纳原则得出的学习机是不可证伪的例子，此处略去。)。这样，我们就可以根据函数集的 VC 维判断学习过程的收敛性。

6.5　关于学习机推广能力的界

VC 维是一个函数集所具有的特性，它直接影响一个学习机的推广性能。应用

它还可以构造出可计算的推广能力的界。

6.5 节讨论的定理和结论的详细证明可参看 Vapnik (2004)。

定理 6.2,6.3 和 6.4 给出了学习机(依概率)收敛的条件,我们更关心的是学习机的推广能力的界,即以多大的概率保证学习机得到的结果的偏差不超过某固定值?

6.5.1 基于生长函数的界

我们直接用生长函数 $G^{\Lambda}(l)$ 来讨论推广能力的界,因为它可以保证快速收敛且与概率测度定义无关,略去了关于 VC 熵和退火 VC 熵的讨论。首先,下面列出了几个重要的基本不等式。

定理 6.6 设 $\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$ 为指示函数集,$G^{\Lambda}(l)$ 为其生长函数,则成立下列两个不等式

$$p\left\{\sup_{\alpha\in\Lambda}\left|\int Q(\boldsymbol{z},\boldsymbol{\alpha})\mathrm{d}F(\boldsymbol{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\boldsymbol{z}_i,\boldsymbol{\alpha})\right|>\varepsilon\right\}\leqslant 4\exp\left[\left(\frac{G^{\Lambda}(2l)}{l}-\varepsilon^2\right)l\right],$$

$$p\left\{\sup_{\alpha\in\Lambda}\frac{\left|\int Q(\boldsymbol{z},\boldsymbol{\alpha})\mathrm{d}F(\boldsymbol{z})-\frac{1}{l}\sum_{i=1}^{l}Q(\boldsymbol{z}_i,\boldsymbol{\alpha})\right|}{\sqrt{\int Q(\boldsymbol{z},\boldsymbol{\alpha})\mathrm{d}F(\boldsymbol{z})}}>\varepsilon\right\}\leqslant 4\exp\left[\left(\frac{G^{\Lambda}(2l)}{l}-\frac{\varepsilon^2}{4}\right)l\right]。$$

引入前面的简化记号 $R(\boldsymbol{\alpha})$ 和 $R_{\mathrm{emp}}(\boldsymbol{\alpha})$,上面两不等式变为

$$p\{\sup_{\alpha\in\Lambda}|R(\boldsymbol{\alpha})-R_{\mathrm{emp}}(\boldsymbol{\alpha})|>\varepsilon\}\leqslant 4\exp\left[\left(\frac{G^{\Lambda}(2l)}{l}-\varepsilon^2\right)l\right] \tag{6.15}$$

$$p\left\{\sup_{\alpha\in\Lambda}\frac{|R(\boldsymbol{\alpha})-R_{\mathrm{emp}}(\boldsymbol{\alpha})|}{\sqrt{R(\boldsymbol{\alpha})}}>\varepsilon\right\}\leqslant 4\exp\left[\left(\frac{G^{\Lambda}(2l)}{l}-\frac{\varepsilon^2}{4}\right)l\right] \tag{6.16}$$

定理 6.7 设 $\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$ 为有界实函数集,$A\leqslant Q(\boldsymbol{z},\boldsymbol{\alpha})\leqslant B$,$G^{\Lambda,\mathcal{B}}(l)$ 为其生长函数,则成立下列不等式

$$p\{\sup_{\alpha\in\Lambda}|R(\boldsymbol{\alpha})-R_{\mathrm{emp}}(\boldsymbol{\alpha})|>\varepsilon\}\leqslant 4\exp\left[\left(\frac{G^{\Lambda,\mathcal{B}}(2l)}{l}-\frac{\varepsilon^2}{(B-A)^2}\right)l\right] \tag{6.17}$$

定理 6.8 设 $\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$ 为有界非负实函数集 $0\leqslant Q(\boldsymbol{z},\boldsymbol{\alpha})\leqslant B$,$G^{\Lambda,\mathcal{B}}(l)$ 为其生长函数,则成立下列不等式

$$p\left\{\sup_{\alpha\in\Lambda}\frac{|R(\boldsymbol{\alpha})-R_{\mathrm{emp}}(\boldsymbol{\alpha})|}{\sqrt{R(\boldsymbol{\alpha})}}>\varepsilon\right\}\leqslant 4\exp\left[\left(\frac{G^{\Lambda,\mathcal{B}}(2l)}{l}-\frac{\varepsilon^2}{4B}\right)l\right] \tag{6.18}$$

定理 6.6 和 6.7 是定理 6.5 的推广。

如果说式(6.15)和(6.17)左边是偏差的概率,则式(6.16)和(6.18)左边类似是偏差的均值的概率。

其中 ε 是给定的任意小正数。这些不等式给出的是“最坏(偏差最大)”情况发生的概率的上界。在一致收敛的情况下,对于任意小的 ε,不等式的右端当观测数目趋于无穷时都趋于零。

利用上述不等式可以建立不同学习机推广能力的界。当然,得到的界还是非构造性的。

为了简化后面 SVM 中的表示式,我们引入记号

$$\mathscr{E}=4\frac{G^{\Lambda,\mathscr{A}}(2l)-\ln(\eta/4)}{l} \tag{6.19}$$

定理 6.9　设$\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$为有界实函数集,$A\leqslant Q(\boldsymbol{z},\boldsymbol{\alpha})\leqslant B$。则下不等式以至少$(1-\eta)$的概率同时对$\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$的所有函数成立

$$R(\boldsymbol{\alpha})\leqslant R_{\mathrm{emp}}(\boldsymbol{\alpha})+\frac{(B-A)\sqrt{\mathscr{E}}}{2} \tag{6.20}$$

定理 6.10　设$\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$为有界非负实函数集,$0\leqslant Q(\boldsymbol{z},\boldsymbol{\alpha})\leqslant B$。则下面不等式以至少$(1-\eta)$的概率同时对$\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$的所有函数成立

$$R(\boldsymbol{\alpha})\leqslant R_{\mathrm{emp}}(\boldsymbol{\alpha})+\frac{B\mathscr{E}}{2}\left(1+\sqrt{1+\frac{4R_{\mathrm{emp}}(\boldsymbol{\alpha})}{B\mathscr{E}}}\right) \tag{6.21}$$

这两个不等式限定了函数集合中所有函数的风险上界,当然也包括了最小化经验风险得到的学习机函数。

6.5.2　基于 VC 维的界

利用 VC 维可以得到构造性的且与概率分布无关的界。

考虑具有有限 VC 维 h 的函数集满足式(6.13)

$$G^{\Lambda}(l)\leqslant h\left(\ln\frac{l}{h}+1\right)\qquad(l>h),$$

则式(6.19)可以改写为

$$\mathscr{E}=4\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln(\eta/4)}{l} \tag{6.22}$$

于是就得到了构造性的界。

定理 6.11　设$\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$为有界实函数集,$A\leqslant Q(\boldsymbol{z},\boldsymbol{\alpha})\leqslant B$,则以至少$(1-\eta)$的概率同时对$\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$的所有函数成立不等式

$$R(\boldsymbol{\alpha})\leqslant R_{\mathrm{emp}}(\boldsymbol{\alpha})+\frac{(B-A)\sqrt{\mathscr{E}}}{2} \tag{6.23}$$

定理 6.12 设$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$为有界非负实函数集，$0\leqslant Q(\boldsymbol{z},\boldsymbol{\alpha})\leqslant B$，则以至少$(1-\eta)$的概率同时对$\{Q(\boldsymbol{z},\boldsymbol{\alpha})|\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$的所有函数成立不等式

$$R(\boldsymbol{\alpha})\leqslant R_{\text{emp}}(\boldsymbol{\alpha})+\frac{B\mathscr{E}}{2}\left(1+\sqrt{1+\frac{4R_{\text{emp}}(\boldsymbol{\alpha})}{B\mathscr{E}}}\right) \tag{6.24}$$

这些界均是可以计算的。如对于第 2、第 3 章的 SVM 分类，依据这两个定理，以概率$(1-\eta)$我们可以断定：测试样本不能被最优划分超平面正确分类的界是

$$P_{\text{error}}\leqslant\frac{m}{l}+\frac{\mathscr{E}}{2}\left(1+\sqrt{1+\frac{4m}{l\mathscr{E}}}\right) \tag{6.25}$$

其中，m 为没有被最大间隔分类超平面正确分类的训练样本数，l 为训练样本数，h 为最大间隔分类超平面集合的 VC 维。

若 $m=0$，即分类的经验误差为零(没有分类错误)，则以概率$(1-\eta)$可以断定：测试样本不能被最优划分超平面正确分类的界变为

$$P_{\text{error}}\leqslant\mathscr{E}=4\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln(\eta/4)}{l} \tag{6.26}$$

只要注意到在 SVM 分类时 $A=0$，$B=1$ 和 $R_{\text{emp}}(\boldsymbol{\alpha})=0$ 即可从式(6.22)和(6.24)中得到式(6.26)。

这个界是可以构造算法的，且有清楚的物理含义：学习机推广能力的界依赖于样本个数与学习机备选函数集的 VC 维之比(l/h)，该比值越大(即小的 VC 维 h 和大的样本数 l)，推广能力越好。

6.5.3 SVM 方法中 VC 维的求法

现在讨论如何计算 SVM 方法中函数集的 VC 维。在那里，备选函数集是线性划分超平面的集合。

虽然没有计算 VC 维的一般方法，但例 6.6 指出了 n 维线性超平面集合的 VC 维数为$(n+1)$，然而这一结果不能直接应用到 SVM 方法中。因为在 SVM 方法中是通过升维把样本空间映射到一个特征空间中，在特征空间中寻求最优线性划分超平面，而特征空间往往是高维甚至是无穷维的。直接套用例 6.6 的结果将导致很坏的结果。

事实上，在 SVM 方法中我们实际用到的函数集并不是特征空间中所有线性超平面的集合，而是它的一个满足一定条件的子集，即是“Δ-间隔分类超平面”的集合。当空间的维数很高时，这样的特殊函数集的 VC 维一般远远小于$(n+1)$，且是可以估计的。下面给出一个相关定理，这一定理实际上是 1962 年 Novikoff 作为感知器在研究神经网络时提出的一个重要定理的直接推论。

定理 6.13　设样本 $z\in Z$ 都被包含在一个半径为 R 的球中，即 $\|z\|\leqslant R$，Δ 为一实数，则所有使得样本能够以 Δ-间隔正确分开的分类超平面集合 F 的 VC 维 h 满足估计式

$$h\leqslant \min\left(\left[\frac{R^2}{\Delta^2}\right],n\right)+1 \tag{6.27}$$

其中，中括号意为取整数函数。可以看出，h 不会超过$(n+1)$。当 Δ 很大时，R/Δ 很小，函数集 F 的 VC 维会远远小于$(n+1)$。在应用中常会遇到特征空间维数很高而 Δ-间隔分类超平面集合的 VC 维很小的例子。比如样本向量维数为 256，应用 7 阶多项式核，特征空间维数是 10^{19}，而最大间隔分类超平面集合的 VC 维只有 10^3。通过改变 Δ 可以调整 VC 维，Δ 越大 VC 维越小，这是 SRM 归纳原则的重要依据之一。

这一定理给出实际计算一类函数集 VC 维上界的一个可构造的具体算法，这也正是 SVM 方法中估计 VC 维所依据的方法。

6.6　SRM 归纳原则

前面已经提到，SVM 方法是一种小样本的训练学习方法。大小总是相对的，这里所说的小样本是与函数集的 VC 维数相比。设训练样本数目为 l，函数集的 VC 维数为 h，如果 l/h 较小(比如小于 20)，则认为这样的训练样本集是小样本。

作为适用于大样本数据的 ERM 归纳原则的发展，这里介绍一种适用于小样本数据的新的归纳原则。

从不等式(6.24)可以看出，要最小化实际风险 $R(\boldsymbol{\alpha})$，不但要最小化不等式右边的第一项[(经验风险项 $R_{\text{emp}}(\boldsymbol{\alpha})$]，还要同时考虑不等式右边的第二项(我们称其为置信区间)。不等式(6.24)右边的第一项的大小取决于选取函数集中的某一特定函数，而第二项则取决于整个函数集的 VC 维。因此，要同时最小化不等式(6.24)右边的两项，必须使 VC 维是可控制的。

一般情况下，对于同类的函数，函数集越“大”，VC 维可能越大，函数的复杂性也加大。即，存在

$$\text{若 } A\subset B,\text{则 VC-dim}(A)\leqslant \text{VC-dim}(B)$$

设机器学习的备选函数集 $S=\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}\}$ 具有一种结构，这种结构是由一系列嵌套的函数子集 $S_i=\{Q(\boldsymbol{z},\boldsymbol{\alpha})\mid\boldsymbol{\alpha}\in\boldsymbol{\Lambda}_i\}$ 组成(如图 6.9)。

若这个子集结构满足：

(1) $S_1\subset S_2\subset\cdots\subset S_n\cdots$

(2) 设函数集 S_k 的 VC 维为 h_k，有 $h_1\leqslant h_2\leqslant\cdots\leqslant h_n\leqslant\cdots$

(3) 函数集 S_i 是完全有界的：$0\leqslant Q(\boldsymbol{z},\boldsymbol{\alpha})\leqslant B,\boldsymbol{\alpha}\in\boldsymbol{\Lambda}_i$

则称函数集具有一个容许结构。

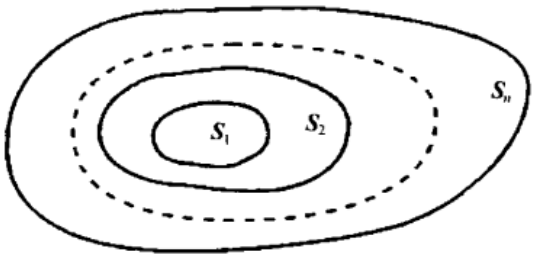


图 6.9 嵌套的子集结构

在有容许结构的函数集上，不等式(6.24)给出的实际风险 $R(\boldsymbol{\alpha})$ 与经验风险和置信区间的关系如图 6.10 所示。

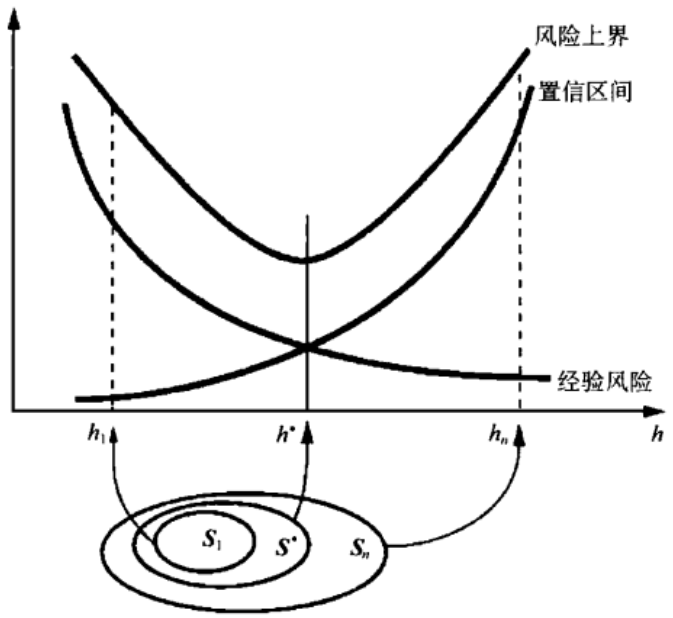


图 6.10 在容许结构函数集中实际风险与经验风险、置信区间的关系

从图 6.10 中可以看出，实际风险的上界函数是经验风险和置信区间两部分的叠加。随着备选函数集的扩大(h 加大)，经验风险减小；同时函数集的 VC 维变大，从而带来置信区间风险的加大。这导致了 SRM 归纳原则。

SRM 归纳原则：在有如上所述的容许结构的函数集中，对一个给定的训练样本集寻求最优函数所依据的原则是——在使风险上界最小的子集 S_k 中，选择使得经验风险达到最小的函数 $Q(\boldsymbol{z},\boldsymbol{\alpha}_l^k)$，从而得到最小的实际风险。

一般地说，随着函数集 VC 维的加大，备选的逼近函数的复杂性加大，而对给定数据逼近的精度提高，同时置信区间变大。SRM 归纳原则是一种确定对给定数据逼近的精度和逼近函数的复杂性间的“折中”方案。

原来的 ERM 归纳原则只是在给定的备选函数集 S 中寻找对数据逼近精度最好的函数；而新的 SRM 归纳原则是通过选择子集，同时考虑逼近精度和置信范围两个

方面。选择子集的标准是使得在这个子集中(如图 6.10 中的 S^*)最小化经验风险泛函,是可以达到实际风险上界的最小值。

SRM 归纳原则是一个一般的原则,它有不同的实现途径。前面五章中介绍的 SVM 方法只是实现这一原则的一种有效途径。

在 SVM 方法中,我们通过改变 Δ-间隔分类超平面的间隔 Δ 来改变函数集的 VC 维,对应不同 VC 维的函数子集形成了一个容许结构,通过寻找最大划分间隔选择子集 S^*,同时找出达到实际风险上界最小值的所求函数(建立的学习机)。

对于 SVM 方法,根据定理 6.12,可以得到实际计算推广能力的界的重要定理。

定理 6.14　如果包含 l 个样本的训练集被最大间隔超平面分开,则测试错误概率的数学期望(对训练集)是以下面 3 个值中最小者的期望为界的

$$EP_{\text{error}} \leqslant E_{\min}\left(\frac{m}{l}, \frac{[R^2 \| w \|^2]}{l}, \frac{n}{l}\right) \tag{6.28}$$

其中 P_{error} 是测试错误概率,l 是训练样本数,n 是训练样本的维数,m 是支持向量的个数,R 是将数据包含其中的超球的半径(若为归一化数据则取 1),$\| w \|^{-1}$ 为间隔值。在任何一种 SVM 的实用算法中,这些量都是能够算出的,因此式(6.28)给出了完全可以计算的学习机推广能力的一个上界。这就是 SVM 学习方法可以对推广能力作出估计的理论根据。不完全令人满意的是,有时这个界有些宽松。在理论上对目前的界的任何改进都有重大意义,当然也是十分困难的。

6.7　本章小结

本章概括介绍了统计学习理论的主要结果,总结如下:

(1) 介绍了传统统计学习中的 ERM 归纳原则,分析了这一归纳原则的不足。

(2) 给出了学习过程一致性收敛的定义、判定的充分必要条件;定义了函数集的 VC 熵、退火 VC 熵和生长函数;给出了号称"机器学习理论三个里程碑"的三个重要定理,它们分别是学习过程是否一致性收敛、是否收敛快,以及收敛是否与概率测度的定义无关的判定依据。

(3) 函数集的 VC 维是决定学习机推广能力的重要概念。通过举例较为详细地介绍了函数集的 VC 维概念,并给出了函数集的 VC 维与生长函数的重要关系。

(4) 讨论了机器学习的推广能力的界——基于生长函数的界是非构造性的,而基于 VC 维的界是可构造性的。

(5) 介绍了 SRM 归纳原则,它是新的统计学习理论的核心问题。

(6) 给出了 SVM 方法中估计推广能力的界的依据。

第 7 章　CMSVM 2.0 软件平台及其使用方法

7.1　CMSVM 软件平台的设计思路

因为 SVM 方法有着坚实的理论基础，所以透彻理解和掌握这一方法需要较多的数学知识。气象业务人员达到这些要求有一定困难。科研人员能否开发出一种软件工具，让想应用该工具解决实际预报问题的人，在对 SVM 方法的基本思想有个初步了解后就能“上手”实际应用呢？

基于这样的思路，我们开发了 CMSVM 这样一个软件平台。应用 CMSVM 做分类和回归预报，用户只需要发挥专业所长做好以下三项工作：

(1) 理清预报思路，确定预报对象及预报因子；

(2) 收集历史样本数据，建立数据文件；

(3) 按要求对数据文件进行归一化(标准化)和格式化处理。

其他建立预报模型和优化模型的工作都可以交给 CMSVM 系统自动完成了。

虽然 SVM 方法具有较好的鲁棒性，但是对于具体的应用问题来说，选取不同的核，或相同核下选取不同的参数都会对预测结果产生明显影响。以 SVM 分类问题为例，对于同一个训练样本集，图 7.1 为选用不同核时的分类结果；图 7.2 为同一核选取不同参数惩罚系数 C 时的分类结果。

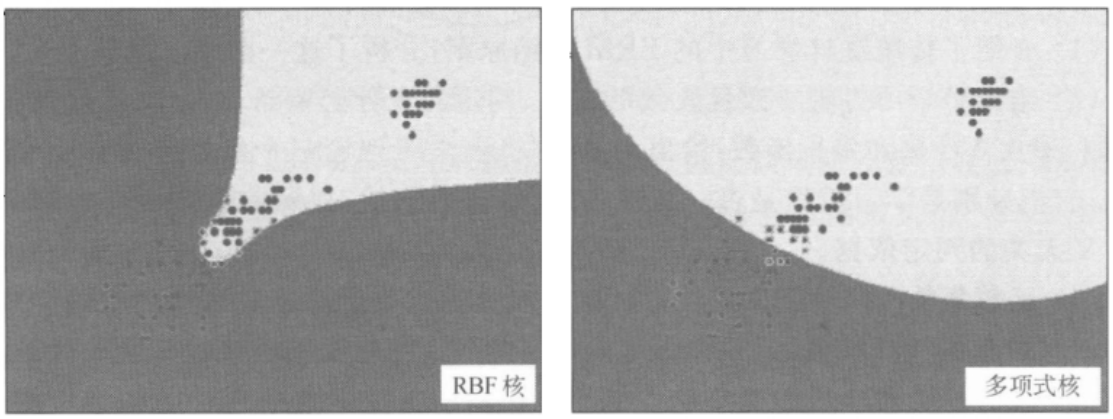


图 7.1　同一数据集不同核的 SVM 方法的分类结果

由图 7.1 和 7.2 可以看出，作为 SVM 方法，每种对给定样本集的分类效果都是很好的，它们都是在特定核和参数条件下计算出来的全局最优解，但它们显然是性能很不一样的分类器。对于一个特定的实际问题来说，只能有一个推广能力最强的解。因此，应用 SVM 方法解决实际问题时，还存在一个要在这些不同核和参数下的最优

解中选优的问题。即使选不出最优的核和参数，也要找出可以实际应用的较优结果。这就是 CMSVM 中的参数优化问题。

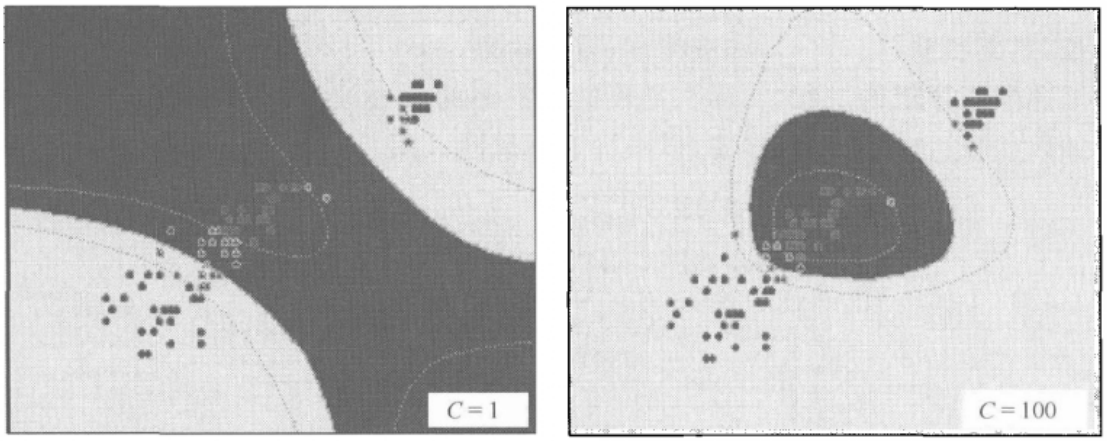


图 7.2　同一数据集同类核不同参数的 SVM 方法的分类结果

非常遗憾的是，至今在理论上对此还没有好的解决方法，这也是目前 SVM 研究中的热点问题之一。

我们认为，对所有应用问题普遍适用的优选参数的方法也许并不存在，人们只能针对不同本质的应用问题采用不同的方法。

从应用看，SVM 的参数优化问题是必须解决的，即使一时不能彻底解决，也要提供实际可用的方法。因为人们总不能随意建立一个模型不问好坏就拿来使用。

通过应用 SVM 方法做天气预报的实践，我们建立了一个软件平台CMSVM 1.0，并提供了一个比较实用的网格自动搜索的参数优化方法。这一参数优化方法的思路和步骤是：

（1）按一定比例对训练样本集进行分割。将原来的样本集 S 分割成三个不相交的子集 S_1，S_2 和 S_3，使 $S=S_1 \cup S_2 \cup S_3$。比如可以使三个集合的样本数大体上的分配比例为 70%，25%和 5%。特别要注意小概率样本在三个集合中的均匀分布。我们分别称 S_1，S_2 和 S_3 为训练集、实验集和检验集。

（2）以样本集 S_1 为训练集，初步选定一组参数 $\boldsymbol{\alpha}$ 进行训练学习，建立学习机 M_o。

（3）以样本集 S_2 为实验集，应用所建立的学习机 M_o 做预报，评估后算得预报效果 T_o。

（4）选定参数 $\boldsymbol{\alpha}$ 的取值增量步长和循环次数，改变参数 $\boldsymbol{\alpha}$，重复步骤(2)和(3)，比较后选出最优预报效果 T^*。

（5）选定最优预报效果 T^* 对应的参数 $\boldsymbol{\alpha}^*$ 为最优参数，由这组参数建立的学习机 M^* 就是经过优化参数后选定的学习机。

（6）应用学习机 M^* 对样本集 S_3 做预报效果的检验。因为样本集 S_3 完全没有

参与学习和参数优化过程，属于独立样本，所以对它的预报效果较为客观地反映了 M^* 的推广能力。

(7) 如果对于步骤(6)的结果不满意，可以在 $\boldsymbol{\alpha}^*$ 附近选择更小的参数区间和增量步长，重复步骤(4)—(6)，加细搜索，直至满意。

上述参数寻优的过程，实际上是根据实验集的预报效果对不同参数下建立的学习机加以比较，选出效果最佳者，以减少实际风险。图 7.3 给出了这一过程的框图。

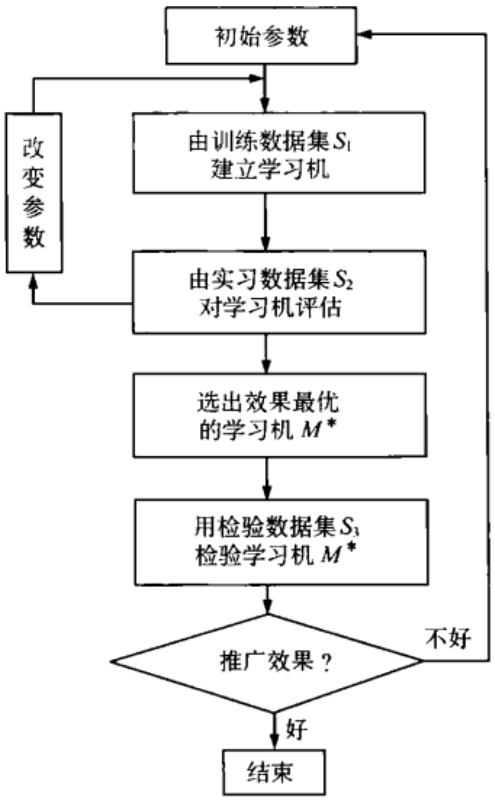


图 7.3 网格搜索优化参数框图

在这一过程中，有几个要点需要加以解释：

(1) 参数优选依据

在步骤(3)中要应用所建立的学习机对实验集 S_2 做预报并评估效果。究竟依据什么方法做评估？在 CMSVM 中对于 SVM 分类和回归提供了不同的评估方法。

SVM 分类是根据正样本的 TS 评分(在 CMSVM 2.0 中增加了负样本的 TS 评分和全体样本的准确率)，选取 TS 评分最高的模型。所谓正样本的 TS 评分是根据分类结果计算得到的一个百分数

$$\mathrm{TS}=\frac{\text{预报正确的正样本数}}{\text{实际正样本数}+\text{空报的正样本数}}\times 100\%。$$

对于 SVM 回归，则可以在均方差、绝对差和相关系数中任选其一(在CMSVM 2.0

中增加了距平)。

(2) 优化参数的网格搜索

SVM 分类中需要调整的参数通常为 2 个,SVM 回归中需要调整的参数通常为 3 个。每个参数都需要给定循环次数、初值和每次循环的增加量步长,每一次循环都是一个完整的学习过程,都要生成一个新的 SVM 并保留所有有用信息。当数据集较大而循环次数较多时,需要大量的运算时间和计算机资源。CMSVM 软件平台对此进行了充分考虑,采取了一系列动态清理存储空间的软件改进方法。即使如此,我们也建议在应用中采取逐步优化的方式,化一次大循环为多次小循环,不要企图经过一个大循环就可以找到理想结果。必须指出,这里给出的参数优化方法是对参数按一定步长搜索的方法,是一种"没有办法的办法"。从理论上说,这种方法只是保证选得能被认可的较优结果,并不能保证是最优。如果经过这种方法总找不到可用的结果,建议对原始样本数据集重新进行审核和分析,看看是否样本序列太短,是否有矛盾数据,还是样本因子的选取有些问题。

(3) TS 评分相同时的处理

在优化 SVM 分类问题的参数时,经常发生多个模型 TS 评分相同的情况。此时必须引入更进一步的评估指标。比如可以采用训练数据集的 TS 评分、训练数据集的 VC 维数等。在 CMSVM 1.0 中采用的是训练数据集的 TS 评分,在对实验数据集的 TS 评分相同的模型中,选取这些模型中对训练数据集的 TS 评分最高的学习机为优者。经过这样的二级评估,结果基本上为唯一。

在 CMSVM 2.0 中我们对参数寻优的方法做了改进,补充了对全体样本集 S 做交叉验证的方法,这一方法更加有效。由于 CMSVM 2.0 包含了 CMSVM 1.0 的全部功能,下面只介绍 CMSVM 2.0。

7.2　CMSVM 2.0 软件平台概述

有关 SVM 的理论和应用研究的国际学术交流近年来一直十分活跃。国际核机器研究组织建立了一个学术交流的网站(http://www.kernel-machines.org),可以直接从其上下载所需的论文、教材和软件,其中提供了多种实现了 SVM 核心算法的程序,如 SVM^{Light},LibSVM,HeroSVM 等。

CMSVM 中采用的是 Thorsten Joachim 的快速 SVM 算法(Joachim T 1999),在保留 SVM^{Light} 内核的基础上,我们进行了深层次的软件开发,形成了集训练学习、参数寻优、模型评估、结果分析和预报等功能为一体的全中文应用软件平台,可用于分类预报和回归预报,特别适合于解决高度非线性的分类和回归等问题。

CMSVM 2.0 除了具有一般 SVM 软件的训练学习的核心功能外,还根据气象业务实际需要增加了如下功能:

(1) 提供了参数循环寻优功能,并可以通过选择不同的寻优方式和标准来选择

最优模型。对于分类预报，可以将正样本 TS 评分、负样本 TS 评分和全体样本的预测精度作为寻优的依据；对于函数估计（回归）问题，可选择绝对差、相关系数、均方差和距平作为寻优的依据。

（2）在软件的运行能力上，增加了循环控制、迭代控制、存储控制等手段，大大提高了程序的运行效率。

（3）增加了贴近度计算和分析模块，可用来在样本空间和特征空间中对训练数据集的品质进行分析（可以检查是否有矛盾数据），进而对模型参数的选取提供参考意见。

（4）增加了两种交叉验证功能。“交叉验证 1”可用来对小的数据集采用“顺序留 1 循环验证”的方法对参数寻优建立模型；“交叉验证 2”可用来对大的数据集按任意比例随机分割、任意确定随机实验次数来进行交叉验证，寻求建立稳定的模型，减少泛化风险。

（5）增加了全部寻优运行过程数据及最终运算结果数据的保存功能，为用户提供了强大的查询模块使其可以对这些数据进行统计分析查询，这对于用户分析理解模型是十分有用的。

（6）提供了便于使用的预报平台，可同时进行多项预报业务。

（7）实现了全中文用户界面，便于业务推广。同时提供了文件管理功能，可在 CMSVM 2.0 系统内实现全部文件管理和编辑功能，使用起来更为方便。

7.3 CMSVM 2.0 的安装及系统目录结构

CMSVM 2.0 在奔腾 I 以上的硬件配置并在中文 Windows 95，98，Me，2000，XP 和 Vista 等系统下都可以正确安装和运行。为了更方便地查询运行结果，Microsoft Office 中的电子表格（Excel）软件是必须预装的。

用户可以在 CMSVM 网站（http://streaml. cma. gav. cn/cmsvm/）免费下载 CMSVM 2.0 安装软件和用户密码生成器。CMSVM 2.0 程序采用自解压方式进行安装，双击安装盘中的图标运行CMSVM 安装程序，就可按照屏幕提示进行余下步骤。安装完毕后在 Windows 桌面上生成一个图标，双击这个图标开始进入 CMSVM 2.0 系统。（注意：首次运行CMSVM 2.0时系统会提供给用户一个 10 位的 ID 号，并询问用户密码。将 ID 号输入到下载的密码生成器，即可获得用户密码。输入正确密码后就能正确运行CMSVM 2.0系统。以后再次运行则无须输入密码了，除非重新安装。）

程序默认安装后的目录结构如图 7.4 所示，在程序运行子目录下有 7 个一级子目录，它们分别对应 CMSVM 2.0 系统的 7 个主选项：两类分类、函数估计（回归分析）、预报应用、贴近度计算、交叉验证 1、交叉验证 2 和帮助说明。

“两类分类”和“函数估计”下分别设有 5 个二级子目录，其中的“训练文件”、“实验文件”和“检验文件”子目录分别由用户存放用于学习建模、优化模型和检验结果的

3 个样本数据文件；分类结果或回归结果子目录存放系统每次运行后自动产生的 4 个文件（统计结果、最优模型、计算结果和检验结果）备查；结果另存子目录用来单独存放上述 4 个文件。

“预报应用”下设有 4 个二级子目录，其中的“数据文件”子目录由用户存放用于做未来预报的样本数据文件；“模型文件”子目录中存放的是经过“两类分类”和“函数估计”运算后得到的最优模型文件，是由系统自动存入的；“预报结果”子目录存放预测计算后自动产生的 2 个文件（预报结果文件和用做预报的模型对应数据的匹配文件）备查；“结果另存”子目录用来单独存放上述 2 个文件。

“贴近度计算”下设有 2 个二级子目录，其中的“计算结果”子目录存放计算后自动产生的 6 个结果文件（均为 Excel 文件格式）备查；“结果另存”子目录用来单独存放上述 6 个文件。

“交叉验证 1”和“交叉验证 2” 的目录结构相同（如图 7.5），分别下设 2 个二级子目录：两类分类和函数估计。每个二级子目录下设有 3 个三级子目录，其中的“验证文件”子目录由用户存放用于做交叉验证建模的样本数据文件；“分类结果”或“回归结果”子目录存放系统每次做交叉验证计算运行后自动产生的 4 个文件（统计结果、最优模型、计算结果和检验结果）备查；“结果另存”子目录用来单独存放上述 4 个文件。

“帮助说明”下设有 7 个帮助说明的文本文件。

为了用户初用时运行方便，所有用来存放用户数据文件的子目录中都预装了一些小例子数据文件供用户参考。

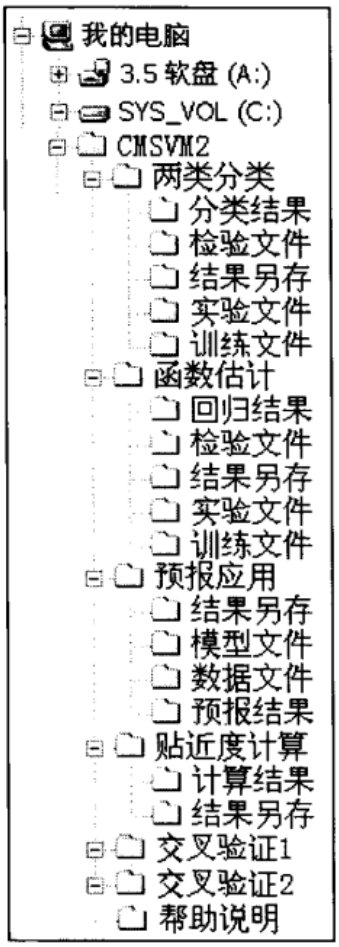


图 7.4　程序目录结构

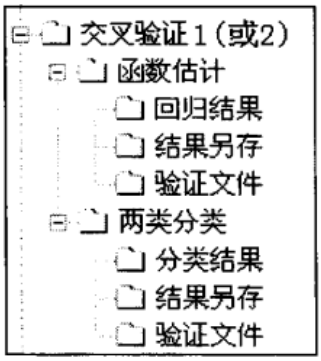


图 7.5　“交叉验证”目录结构

7.4　数据文件格式要求

无论是做两类分类，还是做函数估计、交叉验证或贴近度计算，都需要事先输入数据文件。用户的样本数据要组织成满足一定格式的数据文件才能调入 CMSVM 系统中运行。所有数据文件遵循大体上相同的格式，这种格式要求是为了与国际通行的软件兼容。下面分别加以介绍。

7.4.1 两类分类和函数估计的数据文件格式

用于建立两类分类模型的样本数据格式为：

类别 因子序号:因子值 …# * * * *

例 7.1 已知降水数据如表 7.1：

表 7.1 降水数据表

台站号	日期	降水量	类值	因子 1 值	因子 2 值	…	因子 20 值
53664	03/06/01	0	−1	0.272	0.318	…	0.998
53664	03/06/02	3.5	1	0.661	0.285	…	0.999
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
53664	03/08/21	0	−1	0.174	0.486	…	0.998

则建立的分类数据文件格式如下：

−1	1:0.272	2:0.318	…	20:0.998	#726	0	−1.0	2003	6	1	53664
1	1:0.661	2:0.285	…	20:0.999	#1158	3	1.0	2003	7	7	53664
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1	1:0.174	2:0.486	…	20:0.998	#1698	0	−1.0	2003	8	21	53664

第一列为样本所属类别，在本例中，−1 代表无雨类，1 代表有雨类，这里是两类预报。其余列为因子序号和对应的因子值，需要注意的是因子序号与因子值之间用“:”相连，中间不能有空格；不同因子间为一个(或多个)空格。“#”号后面的内容为注释，可有可无，它们对程序的运行没有影响。本例中注释部分给出的是样本出现的日期等信息，实际运用中可根据自己的需要添加注释。

用于建立函数估计(回归)模型的数据格式与两类分类建模的数据格式雷同，所不同的只是第一列变为预报对象的实况值。

预报对象值 因子序号:因子值…# * * * *

用表 7.1 给出的数据建立的回归数据文件如下：

0.0	1:0.272	2:0.318	…	20:0.998	#726	0	−1.0	2003	6	1	53664
3.5	1:0.661	2:0.285	…	20:0.999	#1158	3	1.0	2003	7	7	53664
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.0	1:0.174	2:0.486	…	20:0.998	#1698	0	−1.0	2003	8	21	53664

有时为了优化所建模型，需要将全部数据样本随机分割为三个部分：用于建模的训练样本子集(约占样本数的 75%)、用于优化模型参数的实验样本子集(约占样本数的 20%)和用于检验模型推广能力的检验样本子集(约占样本数的 5%)。三个数据文件的格式相同，可以先统一做格式处理，而后分割。注意分割后的代表性。

各样本子集的数据整理好以后，必须以文件形式存入各自的文件夹，以便系统程

序自动调用。将分类训练文件存入“两类分类\训练文件”文件夹，将分类实验文件存入“两类分类\实验文件”文件夹，将分类检验文件存入“两类分类\检验文件”文件夹。回归建模与此类似，三个文件分别存入“函数估计\训练文件”、“函数估计\实验文件”和“函数估计\检验文件”文件夹。

7.4.2　预报应用的数据文件格式

预报应用要求的数据文件的格式与“两类分类”和“函数估计”的类似，只是数据文件的第一列全为数值 0，以表示其值待确定。

例 7.2　下面给出的是依据数值预报产品整理出的 0—168 时次的 8 个预报样本数据所形成的预报数据文件：

0	1:0.38272	2:0.40351	3:0.68354	…	20:0.49153	#	56294	2003	9.	1.	0
0	1:0.30864	2:0.56140	3:0.69620	…	20:0.50847	#	56294	2003	9.	1.	24
0	1:0.37037	2:0.45614	3:0.62025	…	20:0.62712	#	56294	2003	9.	1.	48
0	1:0.61728	2:0.57895	3:0.54430	…	20:0.49153	#	56294	2003	9.	1.	72
0	1:0.54321	2:0.75439	3:0.60759	…	20:0.47458	#	56294	2003	9.	1.	96
0	1:0.32099	2:0.36842	3:0.68354	…	20:0.50847	#	56294	2003	9.	1.	120
0	1:0.64198	2:0.68421	3:0.69620	…	20:0.49153	#	56294	2003	9.	1.	144
0	1:0.67901	2:0.61404	3:0.59494	…	20:0.47458	#	56294	2003	9.	1.	168

预报数据集整理好以后，以文件形式存入“预报应用\数据文件”文件夹中。

7.4.3　关于其他数据文件

“贴近度计算”不需单独的数据文件，它是直接调用“两类分类”和“函数估计”中的训练文件子目录下的数据文件。

“交叉验证 1”和“交叉验证 2”都只需要一个数据文件——验证文件，它的格式与两类分类和函数估计中的训练文件的格式相同。

7.4.4　建立数据文件需要注意的几点

(1) 所有建立的数据文件都要求用户自己命名。CMSVM 系统本身对文件的命名并没有特殊要求，中英文名称均可。但是为了应用方便，文件取名时最好有明显的识别特征。

(2) 数据文件可以在 CMSVM 2.0 中的文件编辑器中编辑，也可以在任何编辑软件下编辑，生成后缀为 dat，svm 或 txt 的文件均可。切记：编辑后的文件中不能有空行，尤其是文件最后不要加空行，以免系统错读。

(3) 每一个样本数据是文件中的一个“行”，因此在建立文件行时中间不要加入“硬回车”，以免系统错读。

(4) 数据文件要严格按照上述格式规定建立，否则 CMSVM 系统不读取。

(5) 所有编辑好的数据文件必须存入对应的子目录，否则 CMSVM 系统不能读

取。

(6) 建议所有数据最好按不同因子(依列)单独做归一化处理,这样可以提高运行效率。

7.5 CMSVM 2.0 使用说明

点击桌面上的图标即可进入 CMSVM 2.0 建模预报应用软件平台,出现如图 7.6 所示的主界面。



图 7.6 CMSVM 2.0 主界面

点击界面右上角的按钮,即可退出 CMSVM 2.0 系统。

CMSVM 2.0 主界面提供了 8 个选项(如图 7.7),下面分别加以介绍。

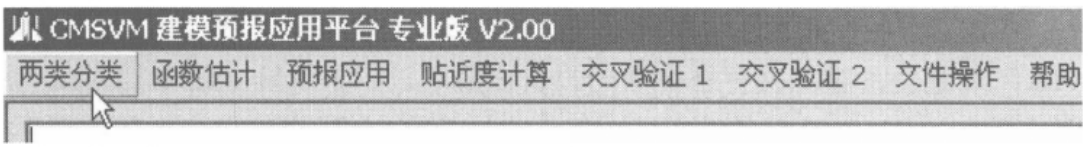


图 7.7 CMSVM 2.0 菜单栏中的 8 个选项

7.5.1 两类分类

CMSVM 2.0 主菜单栏第一个选项为“两类分类”。本选项是用于解决两类分类问题,特别是非线性的两类分类问题。也就是说,依据大量样本数据,通过训练学习,建立分类的决策函数。点击该选项,出现如图 7.8 的下拉菜单,该下拉菜单有 5 个选项,它们分别是“选择文件”、“设置参数”、“计算”、“保存结果”和“查询结果”。这些选项间设置了运行顺序的优先约束,如:没有运行“选择文件”就不可能“设置参数”等。

具体运行方法下面将分别加以介绍。

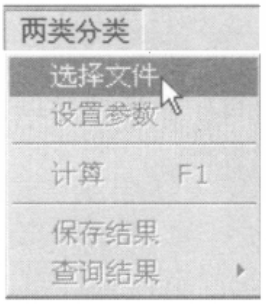


图 7.8　“两类分类”的下拉菜单

(1)“选择文件”。初次安装后只有“选择文件”选项被激活。点击该选项后弹出一个选择文件的窗口(如图 7.9),该界面提供 3 个下拉式文件选择窗口和两个按钮。点击选择窗口右边的箭头,在窗口下出现备选文件名列表,点击所要选择的文件名,该文件即出现在窗口中。依次选择好“训练文件”、“实验文件”和“检验文件”后,点击“确定”按钮自动进入“设置参数”选项的窗口,并把选择结果显示在屏幕的左上方;点击“取消”按钮则放弃所做选择返回主界面。

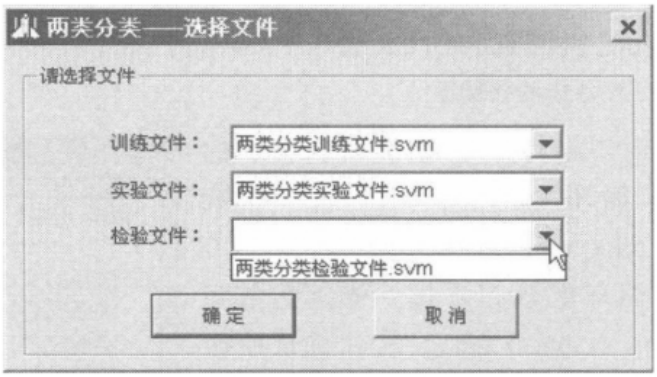


图 7.9　“选择文件”的文件选择窗口

(2)“设置参数”。选择好文件后自动出现“设置参数”的窗口(如图 7.10),该界面提供 2 个下拉式选择窗口、6 个数字窗口和 2 个按钮。

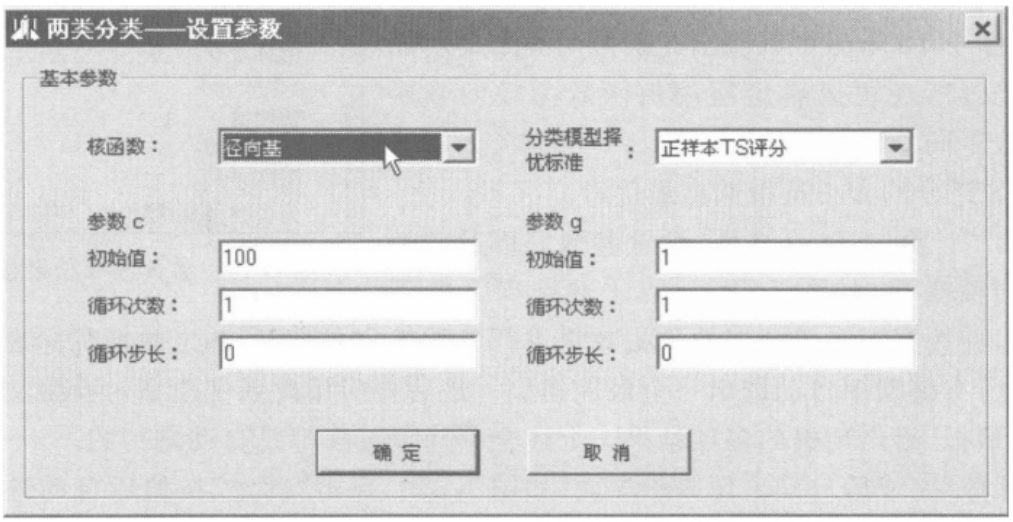


图 7.10　“设置参数”窗口

左面的下拉式选择窗口用来选择核函数,点击选择窗口右边的箭头,出现 8 种备的选核函数列表(如图 7.11),点击所要选择的核函数名称,该核函数即出现在窗口中。

右面的下拉式选择窗口用来选择分类模型的择优方式，点击选择窗口右边的箭头，出现3种备选方案(如图7.12)，点击所要选择的择优方式，该择优方式即出现在窗口中。通常是以正样本的TS评分作为模型的寻优标准，有时也用另外的两种寻优标准。

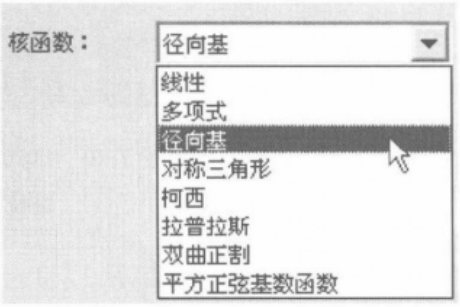


图7.11 备选的核函数

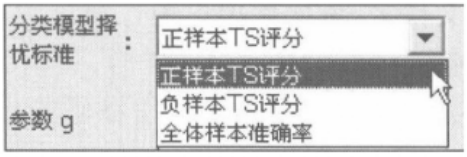


图7.12 分类模型择优标准

左面3个数字窗口是用来确定惩罚系数C的参数寻优的初始值、循环次数和步长；右面3个数字窗口是用来确定对应核函数的参数(如g，u，C等)寻优的初始值、循环次数和步长。依次在数字窗中填好希望的数值后，点击“确定”按钮确认所做选择，选择结果即会显示在屏幕的左方(如图7.13)，而后返回主界面，等待“运算”；点击“取消”按钮则放弃所做的参数选择返回主界面，但原来的文件选择仍保留。

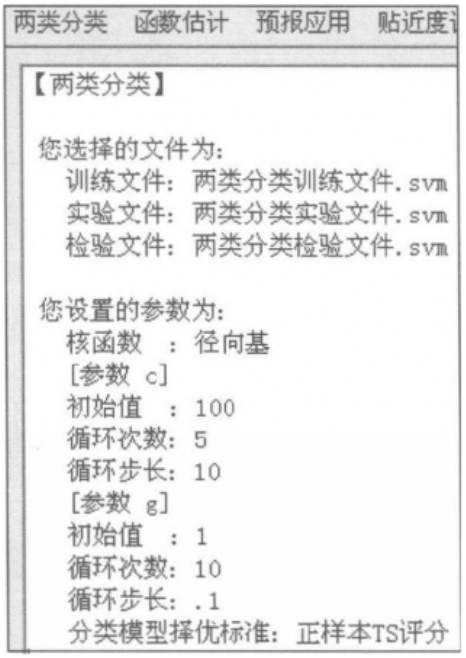


图7.13 参数选择结果显示

从图7.13不但可以看出此次做两类分类所用的数据文件，还可以看出所用核函数为径向基函数，选取不同参数循环建模$5\times10=50$次，依据正样本TS评分的大小从中选出最优者。

(3)“计算”。选择文件、设置参数完成后，“计算”选项被激活(如图7.14)。点击此选项开始训练建模，并按照对实验数据的预报效果进行参数优化，按照所选定的择优标准在循环所建的所有模型中自动选出一个最优模型。最后再应用此最优模型对检验文件的数据做出预报，给出预报的多项结果。上述全部计算过程都是自动进行的。

“计算”完成后自动出现提示界面(如图7.15)，点击“确定”按钮后自动进入查询当前运行结果界面(如图7.16)，它提供4个按钮，可以分别查询当次计算得到的“统计结果”、“最优模型”、“计算结果”和“检验结果”。点击“取消”按钮则放弃当前查询返回主界面。



图 7.14　激活“计算”选项

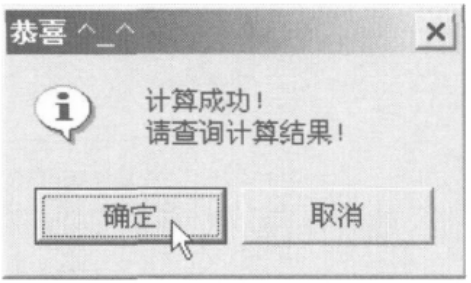


图 7.15　“计算”完成后的提示界面

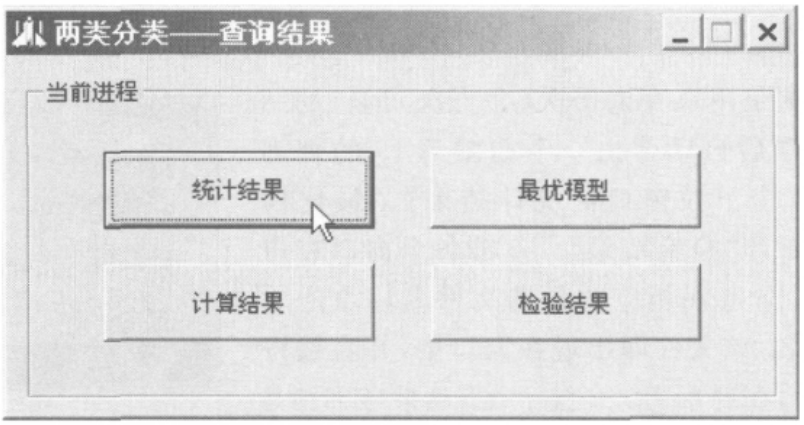


图 7.16　查询结果界面

(4)“保存结果”。每次“计算”完成后,该选项被激活。系统会按规定格式将“计算”后的结果自动统一命名并保存在相应的文件夹中。但如果用相同名称的数据文件重复运算,则只保留最后一次的结果,其余结果均被覆盖。如果想把每一次运行的结果都保存备查,则需要应用“保存结果”选项。点击“保存结果”选项后弹出一个选择保存结果文件的界面(如图 7.17),该界面提供 4 个按钮:“统计结果文件”、“最优模型文件”、“计算结果文件”和“检验结果文件”,分别用来保存 4 类结果文件。点击其中的任意一个按钮,都会弹出类似通常 Windows 系统的文件保存窗口,从中选择希望保存的路径、文件名和保存类型,按“保存”按钮即可完成文件另存并返回主菜单。(注意:“统计结果文件”必须保存成 Excel 后缀为 xls 类型文件,否则不能查询。)

(5)“查询结果”。除初次运行外,该选项总处于激活状态,它为用户提供了多种查询功能。点击该选项后自动下拉出一个下级菜单(如图 7.18),它包含 3 个选项:“当前”、“历史”和“另存”。

“当前”选项是用来查询当次运行的结果,它等效于“计算”后的直接查询。点击“当前”选项后自动弹出一个如图 7.16 的查询界面,它包含 4 个按钮:“统计结果”、“最优模型”、“计算结果”和“检验结果”。点击某一按钮就可以在屏幕上看到相应的查询结果。

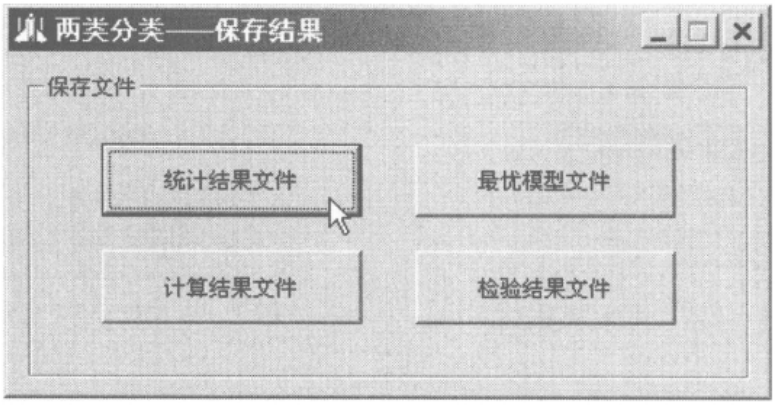


图 7.17 “保存结果”界面

“历史”选项是用来查询历次(含当次)运行的结果。点击该选项后自动弹出一个如图 7.19 的查询界面,它包含 4 个下拉窗口:“统计结果”、“最优模型”、“计算结果”和“检验结果”。点击各个窗口右边的箭头,在窗口下出现相应的备选文件名。点击所要查询的文件名,该文件即出现在窗口中,并直接将该文件内容显示在屏幕上。(注意:“统计结果”显示的是 Excel 文件格式,在屏幕显示的任意数据上双击即可进入 Excel 系统进行相关操作。)



图 7.18 “查询结果”菜单栏

“另存”选项是用来查询用户另存结果文件的选项。点击该选项后自动弹出一个如图 7.20 的查询界面,它包含一个下拉窗口,点击该窗口右边的箭头,在窗口下出现相应的备选文件名列表。点击所要查询的文件名,该文件即出现在窗口中,并直接将该文件内容显示在屏幕上。(注意:如果文件是以 Excel 文件格式另存的,则在屏幕显示的任意数据上双击即可进入 Excel 系统进行操作。)

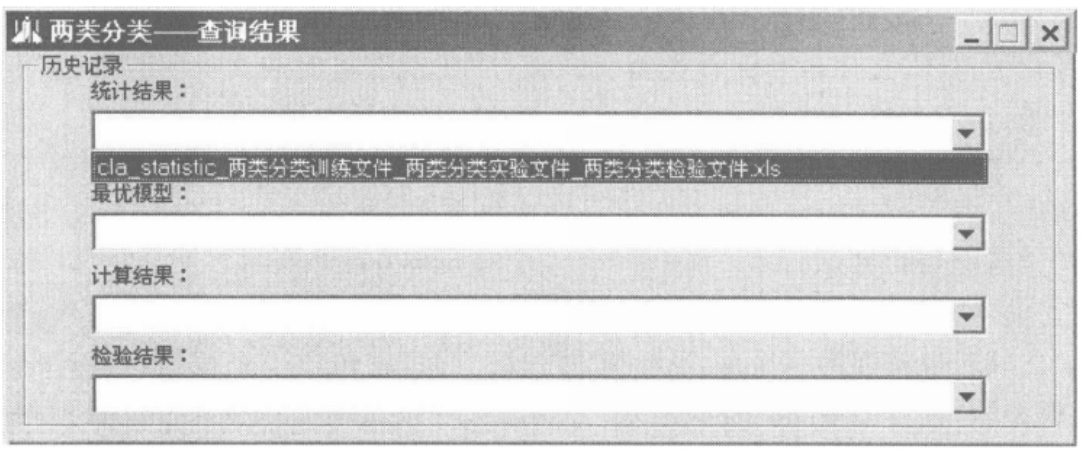


图 7.19 查询“历史”结果界面

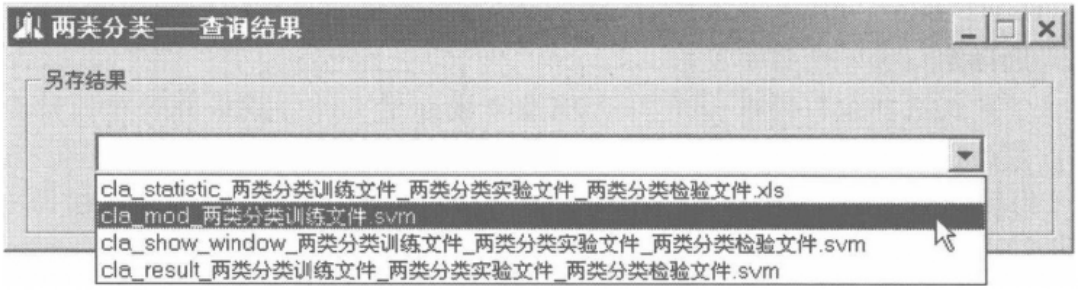


图 7.20　查询“另存”结果界面

7.5.2　函数估计

CMSVM 2.0 主菜单栏第二个选项为“函数估计”，又称回归分析。本选项用来解决回归分析问题，特别是非线性回归问题。也就是说，依据大量样本数据，通过训练学习，求出最优回归函数。点击该选项，出现如图 7.21 的下拉菜单，该下拉菜单与“两类分类”的 5 个选项完全相同。除设置参数外，其他选项的用法与两类分类也完全相同。下面只对“设置参数”做补充介绍。

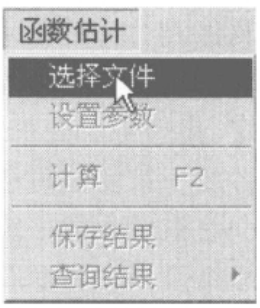


图 7.21　“函数估计”菜单栏

函数估计的“设置参数”界面包括两个区域和两个按钮（如图 7.22）：左半部分是“基本参数”设置区域，右半部分是“其他参数”设置区域，该区域右下方是“确定”和“取消”按钮。

“基本参数”部分用来选取核函数和有关参数，提供了 1 个下拉式选择窗口和 6 个数字窗口，设置方法与“两类分类”相同。

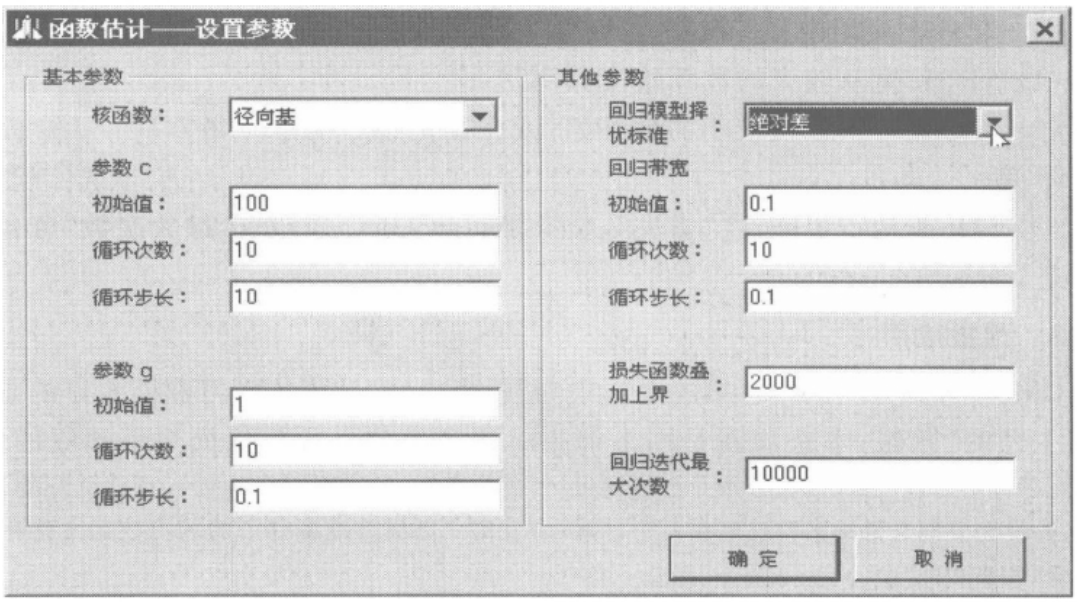


图 7.22　函数估计设置参数

“其他参数”部分用来选取回归(分类)模型择优标准和算法控制参数,提供 1 个下拉式选择窗口和 5 个数字窗口。

该下拉式选择窗口用来选择“回归模型择优标准”,即寻找最优模型所依据的尺度。点击该选择窗口右边的箭头,出现 4 种备选的评估方法列表:“绝对差”、“相关系数”、“均方差”和“距平”。而后,点击所要选择的评估方法名称,该评估方法即出现在窗口中(如图 7.23)。

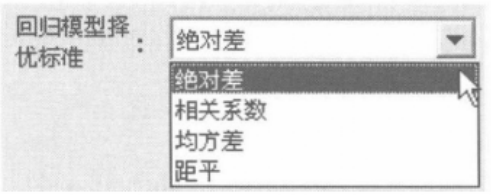


图 7.23 “回归择优标准”窗口

该下拉式选择窗口(如图 7.22)的前 3 个数字窗口是用来确定模型中“回归带宽”参数寻优的初始值、循环次数和步长。“回归带宽”是 SVM 回归中的重要参数,它的大小直接影响建模速度和所建模型的推广能力。一般来说,带宽值越大,支持向量个数越少,回归函数越平坦,但回归累计误差越大。后两个数字窗口是核心算法中的控制参数,前者是用来确定“损失函数叠加上界”,后者是用来确定“回归迭代最大次数”,程序执行过程中若二者之一被超过就会自动终止。建议在应用中根据实际情况适当加大这两个数值。

依次在数字窗口中填好希望的数值后,点击下方的“确定”按钮确认所做选择,选择结果自动显示在屏幕的左方,并返回主界面,等待“运算”;点击“取消”按钮则放弃所做选择返回主界面。需要强调的是,“函数估计”比“两类分类”需要更多的计算时间,而且循环参数也增多了。以图 7.22 选择的参数为例,需要循环建立的模型有 $10\times10\times10=1\,000$ 个,计算量十分巨大。

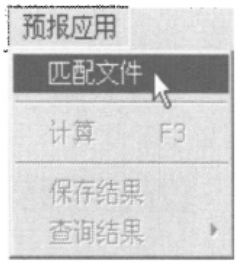


图 7.24 “预报应用”下拉菜单

因此建议在应用中:

(1) 回归带宽从较大的数开始,逐渐变小;

(2) 参数循环次数从较小的数开始,或先固定某一两个参数不变;

(3) 为防止运行中断,将“损失函数叠加上界”和“回归迭代最大次数”可各增大 100 或 1 000 倍。

7.5.3 预报应用

CMSVM 2.0 主菜单栏第三个选项为“预报应用”。训练学习建模的目的是为了应用于预测、预报,本选项就是用来解决应用已经建立的最优模型做预报的问题。(注意:在应用此选项前,模型必须是已经通过“两类分类”或“函数估计”选项计算建立好的最优模型;这些最优模型文件的名称已经被用户掌握,并且用户已经将准备好了的预报数据文件存入“预报应用/数据文件”中。)

点击“预报应用”选项,出现如图 7.24 的下拉菜单,该下拉菜单有 4 个选项,分别是“匹配文件”、“计算”、“保存结果”和“查询结果”。这些选项间设置了运行顺序的优

先约束，如没有运行“选匹配文件”就不可能“计算”等。下面将分别加以介绍。

（1）“匹配文件”。初次应用“预报应用”时只有“匹配文件”选项被激活。点击该选项后弹出一个选择匹配文件的界面（如图 7.25），该界面提供选择“模型文件”和“数据文件”的 2 个下拉式文件选择窗口和 3 个按钮。点击选择窗口右边的箭头，在窗口下出现备选文件名列表，点击所要选择的文件名，该文件即出现在窗口中。依次选择好 1 个“模型文件”和“数据文件”后，点击“确定”按钮后系统自动确认这一对文件匹配，并把匹配结果以一行的形式显示在屏幕的左上方，模型文件在前，数据文件在后。重复这种匹配可以得到多对文件的匹配（同时做多项目预测），并按顺序把所有匹配结果列表在屏幕的左上方（如图 7.26）。点击“完成”按钮后系统自动确认所有的文件匹配，返回主界面，等待“运算”。点击“取消”按钮则放弃所做选择返回主界面。（注意：匹配文件时一定注意模型与数据的对应，切记不要“张冠李戴”！）

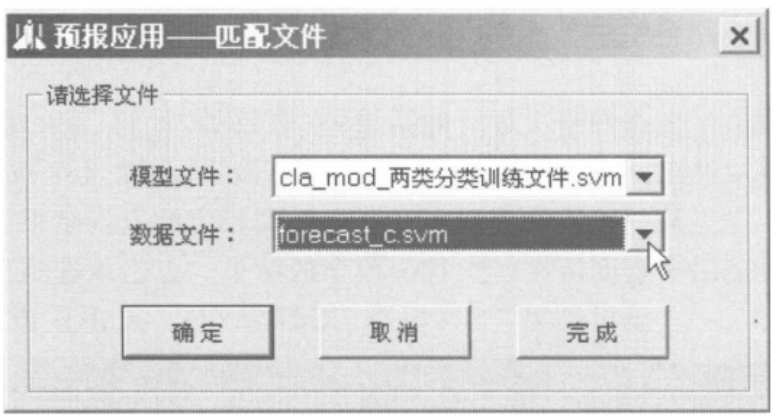


图 7.25　“匹配文件”窗口

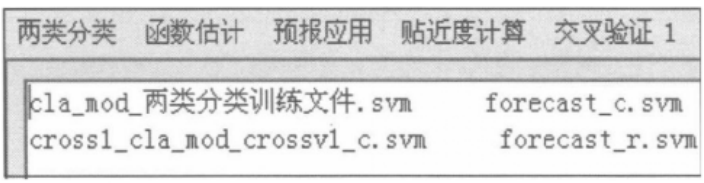


图 7.26　“匹配文件”结果显示

（2）“计算”。用做预测的所有模型文件和数据文件匹配完成后，点击“计算”选项开始预报计算。计算完成后自动出现与图 7.15 相同的提示界面，点击“确定”按钮后自动进入到查询当前运行结果界面（如图 7.27），该界面提供 2 个按钮，可以分别查询“匹配文件”和计算得到的“预报结果”。点击此界面右上角的“×”按钮则放弃当前查询返回主界面。

（3）“保存结果”。每次“计算”完成后，该选项被激活。对于“计算”后的预报结果，系统会按规定格式自动统一命名并自动保存在相应的文件夹中。但如果用相同名称的数据文件重复运算，则只保留最后一次的结果，其余均被覆盖。如果想把每一次运行的预报结果保存备查，则需要应用“保存结果”选项。点击“保存结果”选项后

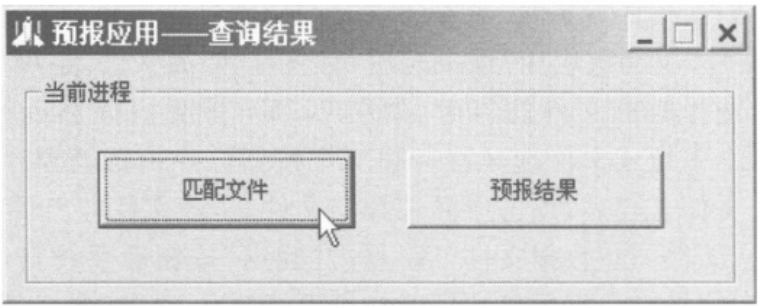


图 7.27 “预报应用”中的“查询结果”界面

弹出类似通常 Windows 系统的文件保存窗口，从中选择希望保存的路径、文件名和保存类型，按“保存”按钮即可完成预报结果文件另存并返回主菜单。

(4)“查询结果”。除初次运行外，该选项总处于激活状态，它为用户提供了多种查询功能。点击该选项后自动下拉出一个下级菜单，它包含 3 个选项：“当前”、“历史”和“另存”。

“当前”选项是用来查询当次运行的结果，它等效于“计算”后的直接查询。点击“当前”选项后自动弹出一个类似图 7.27 的查询界面，它包含 2 个按钮：“匹配文件”和“预报结果”。点击某一按钮就可以在屏幕上看到相应的查询结果。

“历史”选项是用来查询历次(含当次)预报的结果。点击该选项后自动弹出一个查询界面，它包含一个“预报结果”下拉窗口(如图 7.28)。点击该窗口右边的箭头，

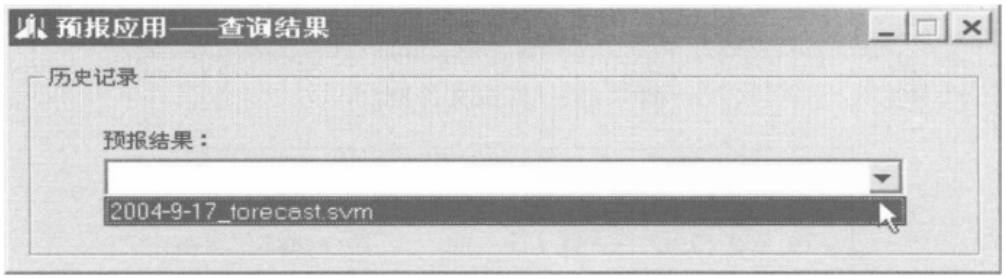


图 7.28 预报应用中的“历史”查询窗口

窗口下出现相应的备选文件名。点击所要选择的文件名，该文件即出现在窗口中，并直接将该文件内容(文件匹配和预报结果)显示在屏幕上。

“另存”选项是用来查询用户另存的预报结果文件。点击该选项后自动弹出一个查询界面，它包含一个类似于图 7.28 的下拉窗口，点击该窗口右边的箭头，在窗口下出现相应的备选文件名。点击所要选择的文件名，该文件即出现在窗口中，并直接将该文件内容显示在屏幕上。

需要注意的是：如果同时做了多项预报(多对文件匹配)，则查询“预报结果”时按文件匹配的先后顺序，依次列出对应的预报结果。原预报数据文件所包含的注释内容也显示在“预报结果”屏幕中。

7.5.4　贴近度计算

CMSVM 2.0 主菜单栏第四个选项为“贴近度计算”。

无论是“两类分类”还是“函数估计”，建模效果的好坏都依赖于训练数据集本身——所选的因子是否与预测输出量关联密切。能否对训练数据集的品质进行分析？进而对模型参数的选取提出参考意见？本选项就是试图解决这些问题的一种尝试。

(1) 什么是贴近度

样本数据是机器学习的基础，没有有效的数据文件就无从建模。训练数据文件的有效性是十分难于检验的一个问题。一种初步然而是合理的方法就是借助贴近度的概念。所谓贴近度，是在训练样本数据集中引入的一种度量函数，它刻画两个训练样本之间的相似或贴近的程度。完全相同的两个样本的贴近度为 1，区别最大的两个样本的贴近度为 0，一般两个训练样本的贴近度介于 0 和 1 之间。

比如，两类分类问题中，正样本和负样本的贴近度应该很小。如果发现正样本和负样本的贴近度很大甚至是为 1，则说明该训练数据集含有不协调甚至是矛盾的数据。如果一个训练数据集的正样本和负样本的贴近度都很小，而正样本和正样本的贴近度及负样本和负样本的贴近度都很大，则该训练数据集是理想的。

对于函数估计问题，也可以做类似的分析。一个很自然的想法，如果两个样本的输出值相差越大，则它们的输入因子间的贴近度应该越小，反之亦然。用户可自己设定输出函数值的两个不重叠的连续区间，落在对应区间的输入因子分别看成是正、负两类样本，这样就可以像两类分类一样做函数估计训练样本的贴近度分析了。

贴近度分析不但可以在样本空间中进行，更重要的是要在经过非线性映射得到的特征空间中进行。在特征空间中，可以借助核函数定义样本的映像间的贴近度。值得提出的是，这一做法给我们提供了一种选取 SVM 核参数的参考途径。

(2) 贴近度如何计算

不但可以考虑样本间的贴近度，而且可以定义样本集与样本集间的贴近度。这些贴近度有多种定义方法，不同定义方法在应用中的效果“各有千秋”。本质上说，贴近度是一种集值测度。

下面给出我们在 CMSVM 2.0 中计算贴近度的公式：

设训练样本向量为 $\boldsymbol{X}_i=(x_{i1},\cdots,x_{in})$ 和 $\boldsymbol{X}_j=(x_{j1},\cdots,x_{jn})$，则它们在样本空间中的贴近度为

$$\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{X}_j)=1-\sqrt{\frac{(x_i-x_j)\cdot(x_i-x_j)}{(x_i+x_j)\cdot(x_i+x_j)}} \tag{7.1}$$

在特征空间中 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 分别变为 $\varphi(\boldsymbol{X}_i)$ 和 $\varphi(\boldsymbol{X}_j)$，则它们在特征空间的贴近度为

$$\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{X}_j)=1-\sqrt{\frac{(\varphi(x_i)-\varphi(x_j))\cdot(\varphi(x_i)-\varphi(x_j))}{(\varphi(x_i)+\varphi(x_j))\cdot(\varphi(x_i)+\varphi(x_j))}} \tag{7.2}$$

将上式中的内积展开，并借助内积与 Mercer 核函数的关系

$$\boldsymbol{K}(\boldsymbol{X}_i,\boldsymbol{X}_j)=(\varphi(\boldsymbol{X}_i)\cdot\varphi(\boldsymbol{X}_j)) \tag{7.3}$$

得到利用核函数来计算的样本 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 在核为 $\boldsymbol{K}$ 的特征空间的贴近度的公式为

$$\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{X}_j)=1-\sqrt{\frac{(\boldsymbol{K}(\boldsymbol{X}_i,\boldsymbol{X}_i)-2\boldsymbol{K}(\boldsymbol{X}_i,\boldsymbol{X}_j))+\boldsymbol{K}(\boldsymbol{X}_j,\boldsymbol{X}_j)}{(\boldsymbol{K}(\boldsymbol{X}_i,\boldsymbol{X}_i)+2\boldsymbol{K}(\boldsymbol{X}_i,\boldsymbol{X}_j))+\boldsymbol{K}(\boldsymbol{X}_j,\boldsymbol{X}_j)}} \tag{7.4}$$

进而可以得到训练样本 $\boldsymbol{X}_i$ 的平均贴近度和方差的计算公式分别为

$$\mathrm{equ}(\boldsymbol{X}_i)=\frac{\sum_{j=1,j\neq i}^{n}\mathrm{sim}(\boldsymbol{X}_i,\boldsymbol{X}_j)}{n-1} \tag{7.5}$$

和

$$\mathrm{eq}(\boldsymbol{X}_i)=\sqrt{\sum_{j=1,j\neq i}^{n}|\boldsymbol{X}_j-\mathrm{equ}(\boldsymbol{X}_i)|^2} \tag{7.6}$$

(3) 如何使用“贴近度计算”选项

“贴近度计算”的下拉菜单有 5 个选项(如图 7.29)，它们分别是“选择文件”、“设置参数”、“计算”、“保存结果”和“查询结果”。这些选项间设置了运行顺序的优先约束，如没有运行“选择文件”就不可能“设置参数”等。下面将分别加以介绍。

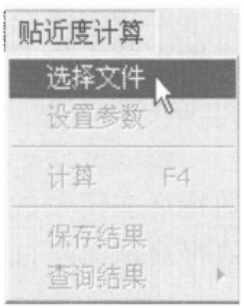


图 7.29 “贴近度”菜单栏

① “选择文件”。初次安装后只有“选择文件”选项是被激活的。点击该选项后弹出一个选择数据文件的界面(如图 7.30)，该界面提供 2 个下拉式文件选择窗口和 2 个按钮。点击选择窗口右边的箭头，在窗口下出现备选文件名列表，点击所要选择的文件名，该文件即出现在窗口中。并且“两类分类”与“函数估计”只能二者择一，点击“确定”按钮自动进入“设置参数”选项的界面(如图 7.31)，并把选择结果显示在屏幕的左上方；点击“取消”按钮则放弃所做选择返回主界面。(注意：做贴近度计算只需要一个数据文件，它自动在“两类分类”和“函数估计”选项中的“训练文件”子目录内全部文件中选取。)

② “设置参数”。选择好数据文件后自动出现如图 7.31 设置参数的界面，该界面提供 1 个下拉式选择窗口、7 个数字窗口和 2 个按钮。

下拉式选择窗口用来选择核函数，点击选择窗口右边的箭头，出现 8 种备选的核函数列表，点击所要选择的核函数名称，该核函数即出现在窗口中。如果选择线性核



图 7.30　进行“贴近度计算”时，“选择文件”窗口

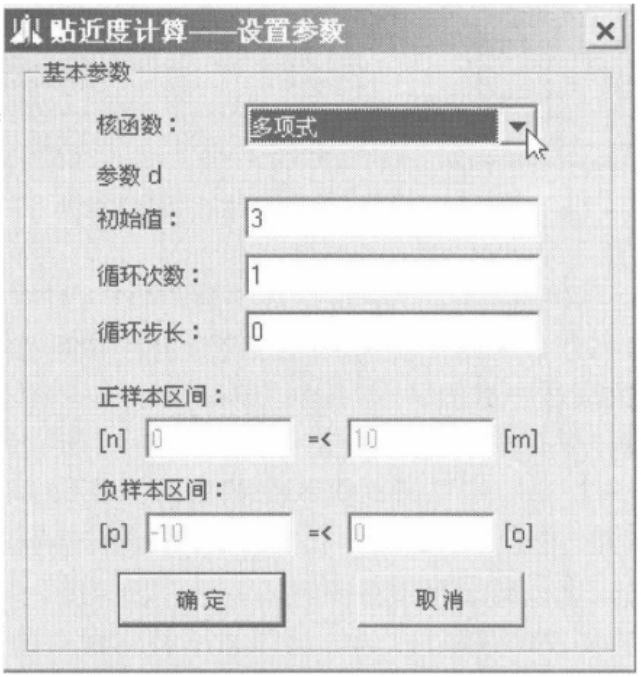


图 7.31　设置“贴近度计算”参数

函数，则是在原始样本空间中做贴近度分析，无须选择其他参数；如果选择其他核函数，则是在该核函数的特征空间中做贴近度分析，此时需要填充前 3 个数字窗口，用来确定对应核函数的参数（如 g，u 和 C 等）寻优的初始值、循环次数和步长。

如果是做函数估计（回归分析）的数据集的贴近度分析，则最后 4 个数字窗口被激活，在其中分别填上对应于正、负样本的输出值区间的左、右端点值。依次在数字窗中填好希望的数值后，点击“确定”按钮确认所做选择，选择结果会显示在屏幕的左方，而后返回主界面，等待“计算”；点击“取消”按钮则放弃所做选择返回主界面。

③ “计算”。选择好数据文件并设置好参数后，点击此键开始贴近度计算，计算出每一正样本和每一正样本的贴近度、每一正样本和每一负样本的贴近度、每一负样本和每一负样本的贴近度，以及相应的最大值、最小值、均值和方差，并对不同的核参

数循环这一过程。计算完成后自动出现熟悉的提示界面，点击“确定”按钮后自动进入查询当前运行结果界面(如图 7.32)，该界面提供 6 个按钮，点击它们可以分别查询计算得到的“正正样本”、“负正样本”、“负负样本”和“正正样本集”、“负正样本集”、“负负样本集”的贴近度计算结果信息。点击该界面右上角的“×”按钮则放弃当前查询返回主界面。

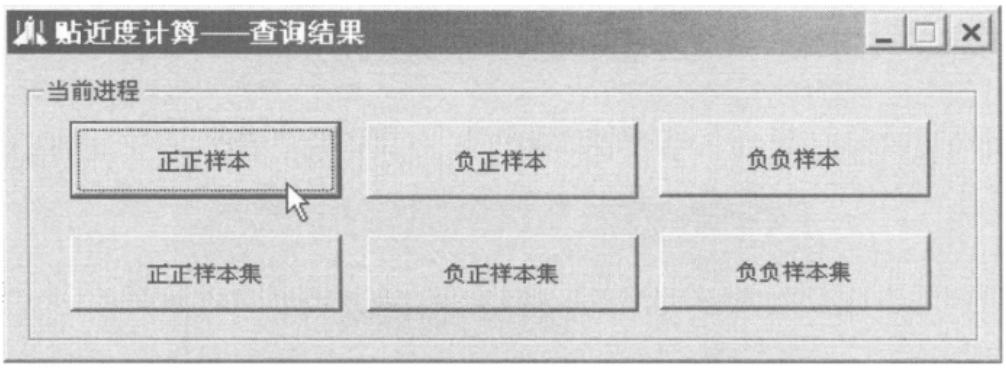


图 7.32 “贴近度计算”的“查询结果”界面

④“保存结果”。每次“计算”完成后，该选项被激活。“计算”后的结果系统会按规定格式自动统一命名并自动保存在相应的文件夹中。但如果用相同名称的数据文件重复运算，则只保留最后一次的结果，其余均被覆盖。如果想把每一次运行的结果保存备查，则需要应用“保存结果”选项。点击“保存结果”选项后弹出一个选择保存结果文件的界面(如图 7.33)，该界面提供 6 个按钮：“正正样本文件”、“负正样本文件”、“负负样本文件”和“正正样本集文件”、“负正样本集文件”和“负负样本集文件”，分别用来保存 6 类贴近度计算结果信息文件。点击其中的任一按钮，都会弹出类似通常 Windows 系统的文件保存窗口，从中选择你所希望保存的路径、文件名和保存类型，按“保存”按钮即可完成文件另存并返回主菜单。(注意：6 类贴近度计算结果信息文件必须保存成 Excel 文件，其后缀 xls 类型，否则不能查询。)

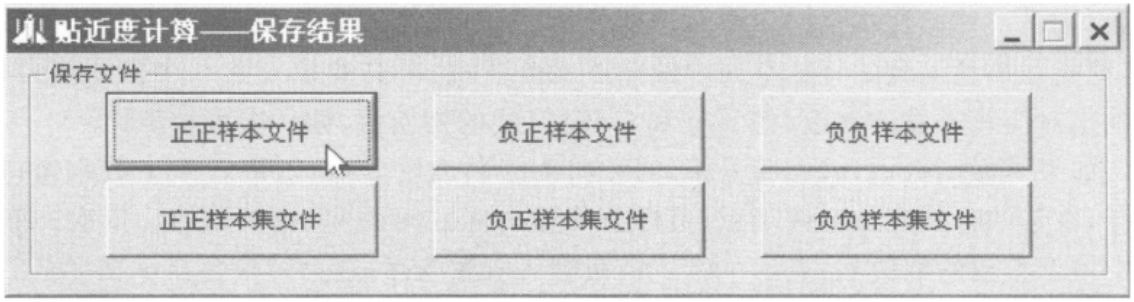


图 7.33 “贴近度计算”的“保存结果”界面

⑤“查询结果”。除初次运行外，该选项总处于激活状态，它提供了查询贴近度计算结果文件的功能。点击该选项后自动下拉出一个下级菜单，它包含 3 个选项：

“当前”、“历史”和“另存”。

“当前”选项是用来查询当次运行的计算结果，它等效于“计算”后的直接查询。点击“当前”选项后自动弹出一个与图 7.32 相同的查询界面，用法也与那里相同，这里不再赘述。

“历史”选项是用来查询历次(含当次)运行的结果。点击该选项后自动弹出一个查询界面(如图 7.34 所示)，它包含 3 个下拉窗口：“正正样本|正正样本集”、“负正样本|负正样本集”、“负负样本|负负样本集”。点击各个窗口右边的箭头，在窗口下出现相应的备选文件名。点击所要选择的文件名，该文件即出现在窗口中，并直接将其内容显示在屏幕上。

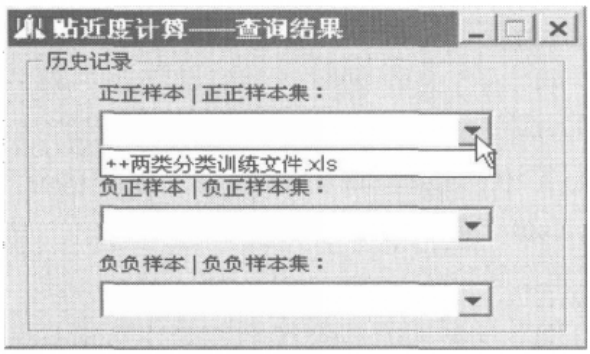


图 7.34　“贴近度”历史结果查询

“另存”选项是用来查询用户另存的所有计算结果文件，该选项只有在做了“另存”操作后才会被激活。点击该选项后自动弹出一个包含下拉窗口的查询界面(如图 7.35)，点击该窗口右边的箭头，在窗口下出现相应的备选文件名。点击选择的文件名，该文件即出现在窗口中，并直接将该文件的内容显示在屏幕上。

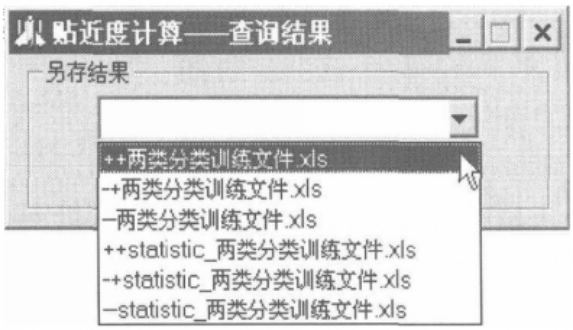


图 7.35　贴近度另存查询

注意：贴近度计算的所有查询结果文件都是 Excel 文件格式，在屏幕显示出来的任意数据上双击即可进入 Excel 系统进行进一步的操作。

7.5.5 交叉验证

无论是两类分类还是函数估计，建模效果的好坏都依赖于训练数据集本身。除了需要考虑所选的因子是否与预测输出量间的关联密切外，还需要考虑训练样本集的代表性。而训练样本集的获得具有一定的随机性，特别是对于小的训练样本集。本选项就是针对这些情况提出了两种解决办法，分别对应于“交叉验证 1”和“交叉验证 2”选项。

(1) 什么是交叉验证？

所谓“交叉验证”就是在总的样本数据集中，按一定比例随机抽取出一部分样本作为训练集，把剩余的样本作为检验集，训练建模后记录对检验集的预报结果。经过多次随机抽取，分析比较所建模型的稳定性。

因此，“交叉验证”可以看作是在数据集中随机抽取一部分数据作为训练文件建模，再用另一部分数据做效果验证。这个过程通过对训练样本的多次随机抽取而“交叉”进行。显然，它相当于多次进行“两类分类”和“函数估计”建模。进行的次数越多，所得到的结果就更具有代表性。

(2) “交叉验证 1”用法

“交叉验证 1”是针对只有很少的训练样本的学习问题设计的——只有几十个样本，无法再把它们划分成“训练样本”、“实验样本”和“检验样本”文件。该选项是先选择一组参数，抽取出一个样本作为待验证的检验样本，用剩余的所有样本建模，顺序对所有样本循环此过程。比如数据集有 50 个样本，则每组参数要循环 50 次。改变一组参数后重复此过程。最后综合对所有样本验证的结果，选取最优参数和最优模型。显然，它是为小的数据集设计的，大的数据集一般不宜使用此选项。

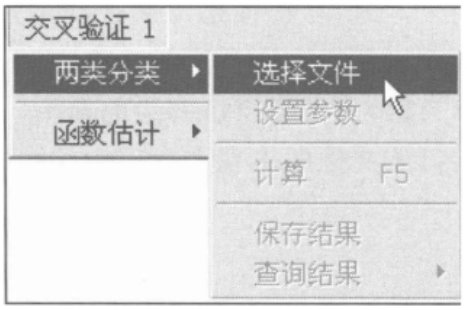


图 7.36 “交叉验证 1”菜单选项

该选项的下拉菜单只有 2 个选项(如图 7.36)，它们分别是“两类分类”和“函数估计”。选择其一之后，弹出的下级菜单与主菜单中的“两类分类”和“函数估计”选项相同，使用方法也基本相同，这里不再重复。不同之处主要在“选择文件”上，这里只需要选取一个数据文件就足够了(如图 7.37)。

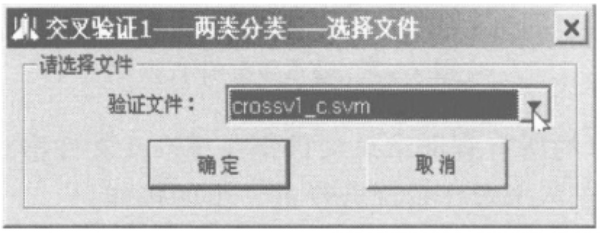


图 7.37 “交叉验证 1”选择文件窗口

(3)“交叉验证 2”用法

对于较大的训练数据集,“交叉验证 1”计算起来非常费时(对于每一组核参数,交叉验证的次数都等于样本数),而且每次只抽出一个样本做验证,缺乏代表性。“交叉验证 2”选项就是针对这两个问题设计的。首先,它可以按任意比例随机把样本数据集切割成训练样本和验证样本,使验证更具有代表性;其次,它可以控制随机实验的次数,以防运算量过大。

该选项的下拉菜单也有 2 个选项(如图 7.36),它们分别是“两类分类”和“函数估计”。选择其一之后,弹出的下级菜单与主菜单中的“两类分类”和“函数估计”选项相同,使用方法也基本相同,这里不再重复。不同之处主要有:

①在“选择文件”上,同“交叉验证 1”,只需选取一个数据文件;

②在“设置参数”上,增加了 2 个数字窗口,分别用来填写用户做交叉验证的“样本比例”和“循环次数”,如图 7.38 和 7.39。这里的“交叉验证样本比例”是指用来做验证的样本占总样本的百分数(如果不是整数按取整函数舍去小数部分),这里的“循环次数”是指随机实验的次数。

建议用户在熟悉了“两类分类”和“函数估计”的使用后再使用“交叉验证 1”和“交叉验证 2”选项。

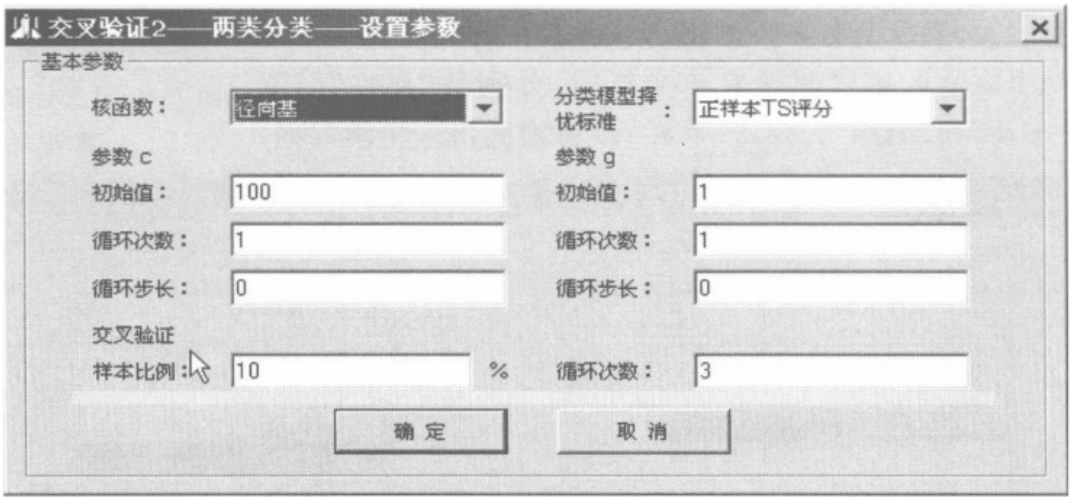


图 7.38　“交叉验证 2”两类分类的参数设置

7.5.6　文件操作

该选项主要是方便用户在 CMSVM 系统内做文件管理和编辑。

“文件操作”的下拉菜单有 2 个选项,分别是“文件管理器”和“文件编辑器”(如图 7.40 所示),下面将分别加以介绍。

(1)“文件管理器”。其用来管理文件,点击该选项后弹出一个文件管理界面(如

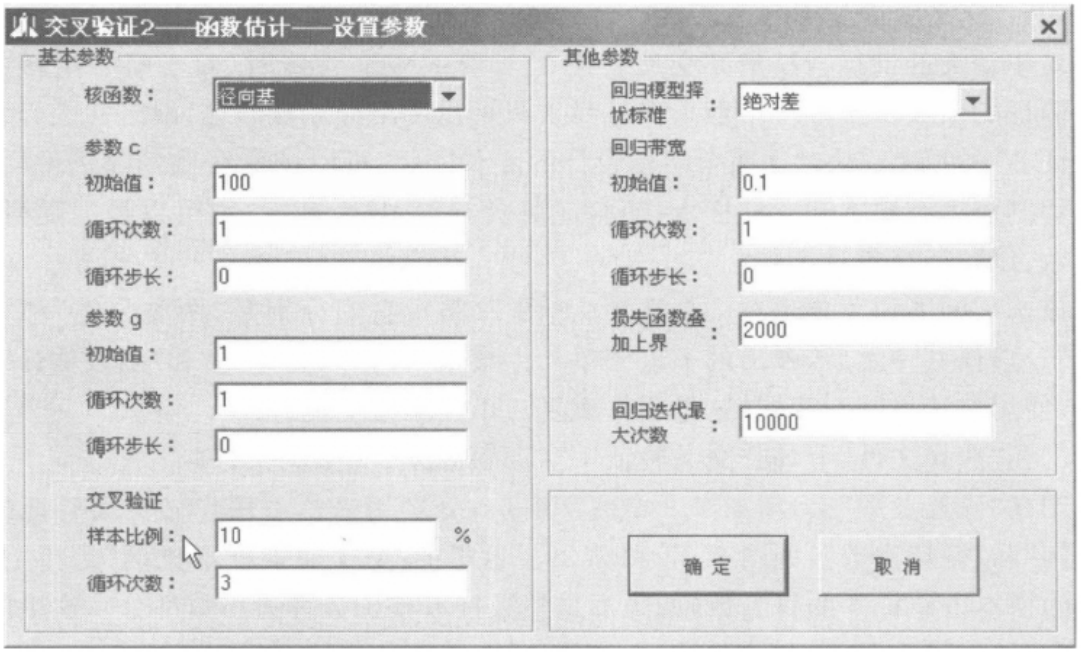


图 7.39 “交叉验证 2”函数估计的参数设置

图 7.41),该界面提供两列文件显示区域:左边是源文件区,右边是目标文件区。每个区域有 3 个下拉式选择窗口,分别用来选择磁盘、磁盘上的文件夹,以及文件夹中的文件。



图 7.40 “文件操作”菜单

两个显示区域之间设有 4 个按钮,分别是“属性”、“复制”、“移动”和“删除”。点击“属性”按钮则弹出一个新界面,

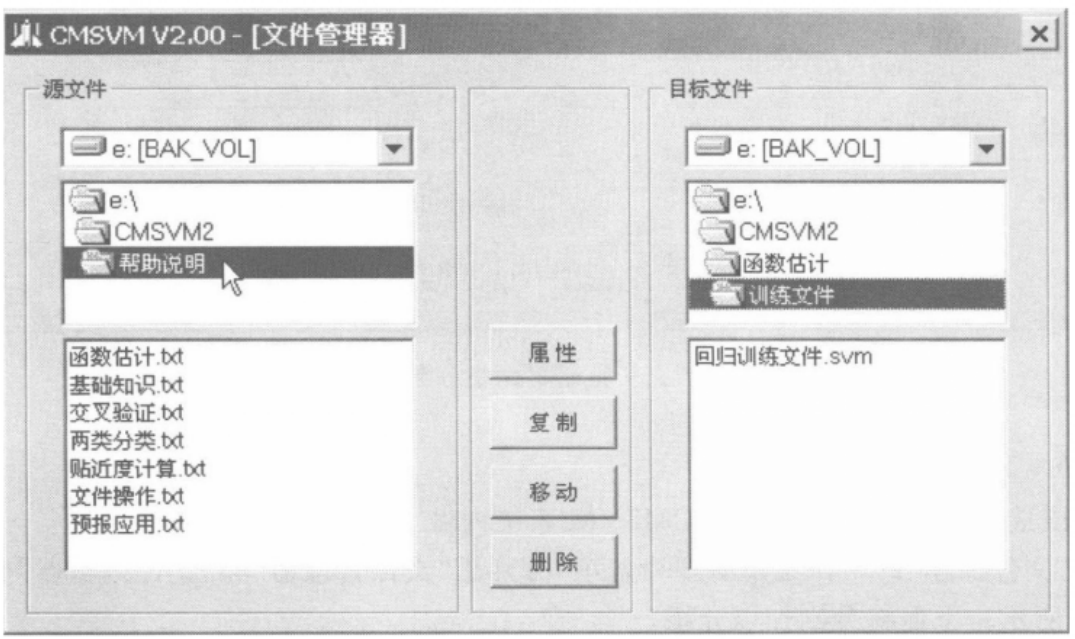


图 7.41 文件管理器

显示左边选中的文件的基本属性；点击“复制”按钮则将左边选中的文件复制到右边选中的目标磁盘的文件夹中；点击“移动”按钮则将左边选中的文件移动到右边选中的目标磁盘的文件夹中；点击“删除”按钮则会将左边选中的源文件删除掉。

(2)“文件编辑器”。其用来建立和编辑文本文件，它是 Windows 文件编辑器的压缩版，无特殊使用要求，这里不再赘述。

无论是两类分类还是函数估计，建模效果的好坏都依赖于训练数据集本身。除了需要考虑所选的因子是否与预测输出量间的关联密切外，还需要考虑训练样本集的代表性。而训练样本集的获得具有一定的随机性，特别是对于小的训练样本集。本选项针对此类情况提出了两种解决办法，即与之对应的“交叉验证 1”和“交叉验证 2”选项。

7.5.7　帮助

该选项为用户提供 CVSVM 2.0 的帮助信息。“帮助”的下拉菜单有 2 个选项，分别是“关于 CMSVM”和“帮助说明”。

点击“关于 CMSVM” 可以查询用户的系统信息和 CMSVM 的版本信息。

点击“帮助说明”出现一个下拉菜单(如图 7.42)，它包含 7 个选项，分别对应于 CMSVM 2.0 的主要功能。点击其中的任一选项，有关的文本信息就显示在屏幕上。

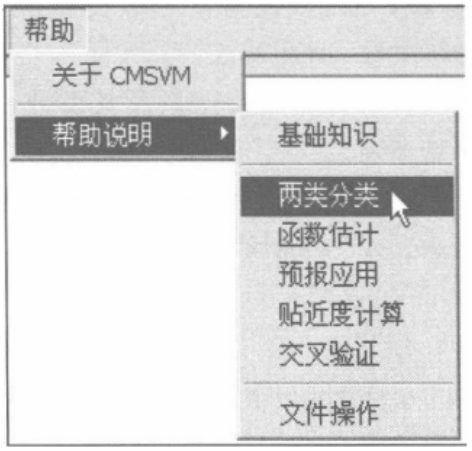


图 7.42　帮助说明

7.6　运行生成文件和对应查询

CMSVM 2.0 系统在每一主选项的每次成功运行后都会自动生成一些相关文件，并自动保存到对应的文件夹中。这些文件都具有固定的命名规则和存储路径，包含一系列重要的统计、查询、分析等信息。CMSVM 2.0 系统每一主选项的“查询结果”主要就是对这些文件的查询。下面将按主选项逐一加以介绍。

7.6.1 “两类分类”的自动生成文件和对应查询

每次运行“两类分类”后，系统自动生成如下 4 个文件：

(1) 统计结果文件

命名规则：cla_statistic_＜训练文件名＞_＜实验文件名＞_＜检验文件名＞. xls

文件类型：Excel 表格

存储文件夹：CMSVM2\两类分类\分类结果

文件内容：该 Excel 表格文件给出模型寻优过程中关于所建立的全部模型的统计信息，刻画了各个样本在参数寻优过程中的“表现”情况。包括每个样本在最优模型中的 α 值、对应的每个样本充当支持向量的次数、每个样本在所有循环中的全部 α 值，以及相关参数的均值、最大值和最小值等，这些信息有助于用户对样本进行分析。

对应查询项目：统计结果查询。

(2) 最优模型文件

命名规则：cla_mod_＜训练文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\两类分类\分类结果

文件内容：该文件为可以做预报的最优模型文件，只能查询不能更改。该文件包含建立最优模型的全部重要参数信息，包括核函数类型、所有模型参数最优值、所有模型参数寻优取值范围、循环次数、循环步长、训练样本个数、样本空间维数、支持向量个数，以及最优模型中系数 b 和所有支持向量及其系数 α 的列表等。

对应查询项目：最优模型查询。

(3) 计算结果文件

命名规则：cla_show_window_＜训练文件名＞_＜实验文件名＞_＜检验文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\两类分类\分类结果

文件内容：该文件给出了参数寻优计算后屏幕显示的全部结果信息。包括进行训练的时间、最优模型参数值；训练样本数据的正、负样本数，正负样本被分类的结果；CPU 运算时间、ε -不敏感损失函数累计值、支持向量个数、VC 维数；对所建最优模型推广能力的误差上界(error)、预测概括率(recall)的下界、预测成功率(precision)的下界、预测 TS 评分下界的估计；最优模型对检验样本的预测结果值(预测正确率、预测概括率、预测成功率、预测 TS 评分)等；此外还给出查询文件输出的信息。

对应查询项目：计算结果查询。

(4) 检验结果文件

命名规则：cla_result_＜训练文件名＞_＜实验文件名＞_＜检验文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\两类分类\分类结果

文件内容：该文件给出了应用所得的最优模型对所选定的“用户检验文件”做预测的结果，列出了样本序号、预测分类值、真实分类值和真实决策函数值。

对应查询项目：检验结果查询。

7.6.2　“函数估计”的自动生成文件和对应查询

每次运行“函数估计”后，系统自动生成如下 4 个文件：

(1) 统计结果文件

命名规则：reg_statistic_＜训练文件名＞_＜实验文件名＞_＜检验文件名＞.xls

文件类型：Excel 电子表格

存储文件夹：CMSVM2\函数估计\回归结果

文件内容：该 Excel 文件给出模型寻优过程中关于所建立的所有模型的统计信息，刻画了各个样本在参数寻优过程中的“表现”情况。包括每个样本在最优模型中的 α 值、对应的每个样本充当支持向量的次数、每个样本在所有循环中的全部 α 值，以及相关参数的均值、最大值和最小值等，这些信息有助于用户对样本进行分析。

对应查询项目：统计结果查询。

(2) 最优模型文件

命名规则：reg_mod_＜训练文件名＞.svm

文件类型：文本文件

存储文件夹：CMSVM2\函数估计\回归结果

文件内容：该文件为可以做预报的最优模型文件，只能查询不能更改。该文件包含建立最优模型的全部重要参数信息，包括核函数类型、所有模型参数最优值、所有模型参数寻优取值范围、循环次数、循环步长、训练样本个数、样本空间维数、支持向量个数，以及最优模型中系数 b 和所有支持向量及其系数 α 的列表等。

对应查询项目：最优模型查询。

(3) 计算结果文件

命名规则：reg_show_window_＜训练文件名＞_＜实验文件名＞_＜检验文件名＞.svm

文件类型：文本文件

存储文件夹：CMSVM2\函数估计\回归结果

文件内容：该文件给出了参数寻优计算后屏幕显示的全部结果信息。包括进行训练的时间，最优模型中的参数值，CPU 运算时间，支持向量个数，ε-不敏感损失函数累计值，用最优模型对检验文件进行回归算得的绝对差、相关系数、均方差和距平等，此外还给出查询文件输出的信息。

对应查询项目：计算结果查询。

(4) 检验结果文件

命名规则:reg_result_<训练文件名>_<实验文件名>_<检验文件名>.svm

文件类型:文本文件

存储文件夹:CMSVM2\函数估计\回归结果

文件内容:该文件给出了应用所得的最优模型对所选定的"用户检验文件"做预测的结果,列出了样本序号、预测值和真实值。

对应查询项目:检验结果查询。

7.6.3 "预报应用"的自动生成文件和对应查询

每次运行"预报应用"后,系统会自动生成一个预报结果文件。

命名规则:做预报的年-月-日_forecast.svm

文件类型:文本文件

存储文件夹:CMSVM2\预报应用\预报结果

文件内容:该文件给出所建模型对预报数据做预测的结果,包括预报模型与数据文件的匹配及预测值。如果是多个待预报样本,则顺序列出预报结果;如果是同时做了多项预报(多对文件匹配),则按文件匹配的先后顺序,依次列出对应的预报结果。此外,预报数据文件的注释内容也包含在"预报结果"文件中。

对应查询项目:预报结果查询。

7.6.4 "贴近度计算"的自动生成文件和对应查询

每次运行"贴近度计算"后,系统自动生成如下 6 个文件:

(1) 正、正样本的贴近度计算结果文件

命名规则:++<训练文件名>.xls

文件类型:Excel 表格

存储文件夹:CMSVM2\贴近度计算\计算结果

文件内容:该 Excel 表格文件给出了用参与循环的第一组参数计算得到的训练数据集中每个正类样本与每个正类样本之间的贴近度列表,并给出了每一正类样本与所有其他正类样本的贴近度值的均值和方差,还给出这些贴近度值的最大和最小值(由于自己和自己的贴近度永远是 1,所以在取最大值时去掉了 1)。

对应查询项目:正正样本查询。

(2) 负、正样本的贴近度计算结果文件

命名规则:-+<训练文件名>.xls

文件类型:Excel 表格

存储文件夹:CMSVM2\贴近度计算\计算结果

文件内容:该 Excel 表格文件给出了用参与循环的第一组参数计算得到的训练数据集中每个负类样本与每个正类样本之间的贴近度列表,并给出了每一负类样本

与所有其他正类样本的贴近度值的均值和方差，还给出这些贴近度值的最大和最小值。

对应查询项目：负正样本查询。

(3) 负、负样本的贴近度计算结果文件

命名规则：－－<训练文件名>. xls

文件类型：Excel 表格

存储文件夹：CMSVM2\贴近度计算\计算结果

文件内容：该 Excel 表格文件给出了用参与循环的第一组参数计算得到的训练数据集中每个负类样本与每个负类样本之间的贴近度列表，并给出了每一负类样本与所有其他负类样本的贴近度值的均值和方差，还给出这些贴近度值的最大和最小值(由于自己和自己的贴近度永远是 1，所以在取最大值时去掉了 1)。

对应查询项目：负负样本查询。

(4) 正、正样本集的贴近度计算统计结果文件

命名规则：＋＋statistic_<训练文件名>. xls

文件类型：Excel 表格

存储文件夹：CMSVM2\贴近度计算\计算结果

文件内容：该 Excel 表格文件给出了参与循环计算的每组参数计算得到的训练数据集中全体正类样本之间的贴近度值的最大和最小值(由于自己和自己的贴近度永远是 1，所以在取最大值时去掉了 1)。

对应查询项目：正正样本集查询。

(5) 负、正样本集的贴近度计算统计结果文件

命名规则：－＋statistic_<训练文件名>. xls

文件类型：Excel 表格

存储文件夹：CMSVM2\贴近度计算\计算结果

文件内容：该 Excel 表格文件给出了参与循环计算的每组参数计算得到的训练数据集中全体负类样本与全体正类样本之间的贴近度值的最大和最小值。

对应查询项目：负正样本集查询。

(6) 负、负样本集的贴近度计算统计结果文件

命名规则：－－statistic_<训练文件名>. xls

文件类型：Excel 表格

存储文件夹：CMSVM2\贴近度计算\计算结果

文件内容：该 Excel 表格文件给出了参与循环计算的每组参数计算得到的训练数据集中全体负类样本之间的贴近度值的最大和最小值(由于自己和自己的贴近度永远是 1，所以在取最大值时去掉了 1)。

对应查询项目：负负样本集查询。

7.6.5 “交叉验证 1”的自动生成文件和对应查询

运行“交叉验证 1”中的“两类分类”和“函数估计”，系统自动生成 4 个不同的文件，下面分别介绍：

(1)“交叉验证 1”中“两类分类”的生成文件和查询

每次运行“交叉验证 1”的“两类分类”后，系统自动生成如下 4 个文件：

① 统计结果文件

命名规则：cross1_cla_statistic_＜训练文件名＞. xls

文件类型：Excel 表格

存储文件夹：CMSVM2\交叉验证 1\两类分类\分类结果

文件内容：该 Excel 电子表格文件给出每个样本在最优模型(这里是用寻得的最优参数对全体样本所建的模型)中的 α 值，以及每个样本在交叉验证寻优过程中充当支持向量的次数。

对应查询项目：统计结果查询。

② 最优模型文件

命名规则：cross1_cla_mod_＜训练文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\交叉验证 1\两类分类\分类结果

文件内容：该文件为可以做预报的最优模型文件，只能查询不能更改，它是由全部训练样本建立的。该文件包含建立最优模型的全部重要参数信息，包括核函数类型、所有模型参数最优值、所有模型参数寻优取值范围、循环次数、循环步长、训练样本个数、样本空间维数、支持向量个数，以及最优模型中系数 b 和所有支持向量及其系数 α 的列表等。

对应查询项目：最优模型查询。

③ 计算结果文件

命名规则：cross1_cla_show_window_＜训练文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\交叉验证 1\两类分类\分类结果

文件内容：该文件给出了参数寻优计算后屏幕显示的全部结果信息。包括进行训练的时间、最优模型参数值；训练样本数据的正、负样本数，正负样本被分类的结果；CPU 运算时间、ε -不敏感损失函数累计值、支持向量个数、VC 维数；对所建最优模型推广能力的误差上界(error)、预测概括率(recall)的下界、预测成功率(precision)的下界、预测 TS 评分下界的估计等；此外还给出查询文件输出的信息。

对应查询项目：计算结果查询。

④ 检验结果文件

命名规则：cross1_cla_result_＜训练文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\交叉验证 1\两类分类\分类结果

文件内容：该文件给出了应用所得的最优模型对全部训练样本做预测的结果，列出了样本序号、预测分类值、真实分类值和真实决策函数值。

对应查询项目：检验结果查询。

(2)“交叉验证 1”中“函数估计”的生成文件和查询

每次运行“交叉验证 1”的“函数估计”后，系统自动生成如下 4 个文件：

① 统计结果文件

命名规则：cross1_reg_statistic_＜训练文件名＞. xls

文件类型：Excel 表格

存储文件夹：CMSVM2\交叉验证 1\函数估计\回归结果

文件内容：该 Excel 表格文件给出每个样本在最优模型(这里是用寻得的最优参数对全体样本所建的模型)中的 α 的值，以及每个样本在交叉验证寻优过程中充当支持向量的次数，这些信息有助于用户对样本分析。

对应查询项目：统计结果查询。

② 最优模型文件

命名规则：cross1_reg_mod_＜训练文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\交叉验证 1\函数估计\回归结果

文件内容：该文件为可以做预报的模型文件，只能查询不能更改。该文件包含建立最优模型的全部重要参数信息，包括核函数类型、所有模型参数最优值、所有模型参数寻优取值范围、循环次数、循环步长、训练样本个数、样本空间维数、支持向量个数，以及最优模型中系数 b 和所有支持向量及其系数 α 的列表等。

对应查询项目：最优模型查询。

③ 计算结果文件

命名规则：cross1_reg_show_window_＜训练文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\交叉验证 1\函数估计\回归结果

文件内容：该文件给出了参数寻优计算后屏幕显示的全部结果信息，包括进行训练的时间、最优模型中的参数值、CPU 运算时间、支持向量个数、ε -不敏感损失函数累计值等，此外还给出查询文件输出的信息。

对应查询项目：计算结果查询。

④ 检验结果文件

命名规则：cross1_reg_result_＜训练文件名＞. svm

文件类型：文本文件

存储文件夹:CMSVM2\交叉验证 1\函数估计\回归结果

文件内容:该文件给出了应用所得的最优模型对全部训练样本做预测的结果,列出了样本序号、预测值和真实值。

对应查询项目:检验结果查询。

7.6.6 "交叉验证 2"的自动生成文件和对应查询

运行"交叉验证 2"的"两类分类"和"函数估计",系统自动生成的文件是不同的,下面分别介绍:

(1)"交叉验证 2"中"两类分类"的生成文件和查询

每次运行"交叉验证 2"的"两类分类"后,系统自动生成如下 4 个文件:

① 统计结果文件

命名规则:cross2_cla_statistic_<训练文件名>.xls

文件类型:Excel 表格

存储文件夹:CMSVM2\交叉验证 2\两类分类\分类结果

文件内容:该 Excel 电子表格文件给出每个样本在最优模型(这里是用寻得的最优参数对全体样本所建的模型)中的 α 值,以及每个样本在交叉验证寻优过程中充当支持向量的次数。

对应查询项目:统计结果查询。

② 最优模型文件

命名规则:cross2_cla_mod_<训练文件名>.svm

文件类型:文本文件

存储文件夹:CMSVM2\交叉验证 2\两类分类\分类结果

文件内容:该文件为可以做预报的最优模型文件,只能查询不能更改,它是由全部训练样本建立的。该文件包含建立最优模型的全部重要参数信息,包括核函数类型、所有模型参数最优值、所有模型参数寻优取值范围、循环次数、循环步长、训练样本个数、样本空间维数、支持向量个数,以及最优模型中系数 b 和所有支持向量及其系数 α 的列表等。

对应查询项目:最优模型查询。

③ 计算结果文件

命名规则:cross2_cla_show_window_<训练文件名>.svm

文件类型:文本文件

存储文件夹:CMSVM2\交叉验证 2\两类分类\分类结果

文件内容:该文件给出了参数寻优计算后屏幕显示的全部结果信息,包括进行训练的时间、最优模型参数值;训练样本数据的正、负样本数,正负样本被分类的结果;CPU 运算时间、ε-不敏感损失函数累计值、支持向量个数、VC 维数;对所建最优模型推广能力的误差上界(error)、预测概括率(recall)的下界、预测成功率(precision)的

下界、预测 TS 评分下界的估计等；此外还给出查询文件输出的信息。

对应查询项目：计算结果查询。

④ 检验结果文件

命名规则：cross2_cla_result_＜训练文件名＞. xls

文件类型：Excel 表格

存储文件夹：CMSVM2\交叉验证 2\两类分类\分类结果

文件内容：该文件顺序列出了应用所得的最优模型参数交叉验证建模后，对每次按给定比例随机抽取的验证样本集所做预测的结果。列出了预测分类值、真实分类值、真实决策函数值和样本序号。

对应查询项目：检验结果查询①。

(2)“交叉验证 2”中“函数估计”的生成文件和查询

每次运行“交叉验证 2”的“函数估计”后，系统自动生成如下 4 个文件：

① 统计结果文件

命名规则：cross2_reg_statistic_＜训练文件名＞. xls

文件类型：Excel 表格

存储文件夹：CMSVM2\交叉验证 2\函数估计\回归结果

文件内容：该 Excel 表格文件给出每个样本在最优参数模型(这里是用寻得的最优参数对全体样本所建的模型)中的 α 值，以及每个样本在交叉验证寻优过程中充当支持向量的次数，这些信息有助于用户对样本进行分析。

对应查询项目：统计结果查询。

② 最优模型文件

命名规则：cross2_reg_mod_＜训练文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\交叉验证 2\函数估计\回归结果

文件内容：该文件为可以做预报的模型文件，只能查询不能更改。该文件包含建立最优模型的全部重要参数信息，包括核函数类型、所有模型参数最优值、所有模型参数寻优取值范围、循环次数、循环步长、训练样本个数、样本空间维数、支持向量个数，以及最优模型中系数 b 和所有支持向量及其系数 α 的列表等。

对应查询项目：最优模型查询。

③ 计算结果文件

命名规则：cross2_reg_show_window_＜训练文件名＞. svm

文件类型：文本文件

存储文件夹：CMSVM2\交叉验证 2\函数估计\回归结果

① 为了方便进一步分析结果，此文件以 Excel 格式保存。

文件内容：该文件给出了参数寻优计算后屏幕显示的全部结果信息，包括进行训练的时间、最优模型中的参数值、CPU 运算时间、支持向量个数、ε-不敏感损失函数累计值等，此外还给出查询文件输出的信息。

对应查询项目：计算结果查询。

④ 检验结果文件

命名规则：cross2_reg_result_＜训练文件名＞.svm

文件类型：Excel 表格

存储文件夹：CMSVM2\交叉验证 2\函数估计\回归结果

文件内容：该文件顺序列出了应用所得的最优模型参数交叉验证建模后，对每次按给定比例随机抽取的验证样本集所做预测的结果。列出了预测值、真实值和样本序号。

对应查询项目：检验结果查询①。

7.7 关于 CMSVM 系统中的核函数

CMSVM 2.0 系统提供了 8 种核函数供用户选用。下面分别给出这些核函数的定义、在程序中核参数的缺省值规定和意义，并通过二元函数的例子给出核函数的直观图像。在核函数的表示公式中，样本表示为 $\boldsymbol{X}=(x_1,\cdots,x_n)$ 和 $\boldsymbol{Y}=(y_1,\cdots,y_n)$，其中 n 为向量的维数。由于线性核就是大家熟悉的由普通点积定义的线性函数，而且没有核参数，所以这里就不单独列出了。

7.7.1 多项式核函数(Polynomial Function Kernel)

$$K(\boldsymbol{X},\boldsymbol{Y})=(s(\boldsymbol{X}\cdot\boldsymbol{Y})+r)^d=\left(s\sum_{i=1}^{n}x_iy_i+r\right)^d \tag{7.7}$$

在 CMSVM 系统中 s 和 r 的值恒为 1，不能改变；d 为可调参数，其默认值是 3，且 d 只取非负整数。

实际应用的多项式核函数都是多元高次多项式，难以直观表示。图 7.43 是一元多项式函数 $y=(x+1)^3$ 的图像。d 的大小决定了曲线的形状。一般来说，d 越大曲线越“曲折”，容易造成“过学习现象”。应用中 d 不宜取值太大，一般不超过 10。

7.7.2 径向基核函数(Radial Basis Function Kernel)

$$K(\boldsymbol{X},\boldsymbol{Y})=\mathrm{e}^{-g\sum_{i=1}^{n}(x_i-y_i)^2} \tag{7.8}$$

其中，g 为可调参数，其默认值是 3，且 g 只取正实数。

① 为了方便进一步分析处理结果，此文件以 Excel 格式保存。

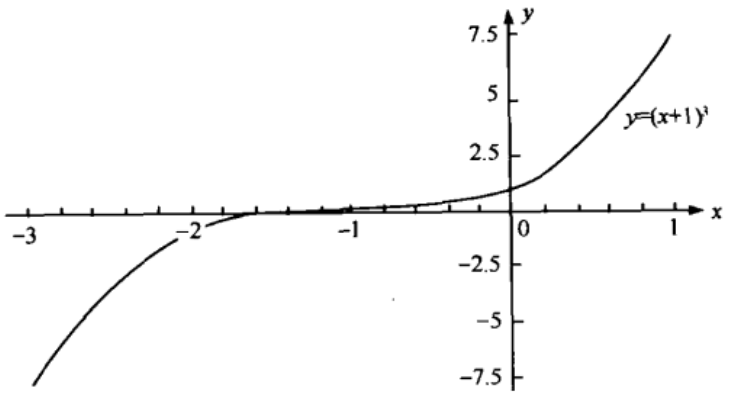


图 7.43　多项式核函数图示

径向基核(RBF 核)，又称高斯核，因为它是由高斯函数(Gauss Function)生成的。图 7.44 是一元高斯函数 $y=\exp(-0.001x^2)$的图像。g 的大小决定了图像的“胖瘦”，g 越大，图像越“尖”。

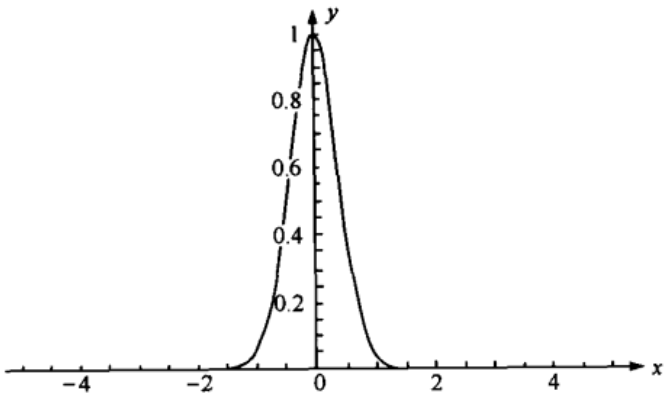


图 7.44　径向基核函数图示

7.7.3　对称三角形核函数(Symmetric Triangle Function Kernel)

$$K(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n} \max(1 - u \mid x_i - y_i \mid, 0) \tag{7.9}$$

u 为可调参数，其默认值是 3，且 u 只取正实数。

图 7.45 给出函数 $y=\max(1-|x|/3, 0)$的图像，是一个等腰三角形。u 的大小决定了图像底边的长短，$2/u$ 为底边的长度。u 越大，拟合误差越小，但容易造成“过学习现象”，影响推广能力。

7.7.4　柯西核函数(Cauchy Function Kernel)

$$K(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{1}{1 + u(x_i - y_i)^2} \tag{7.10}$$

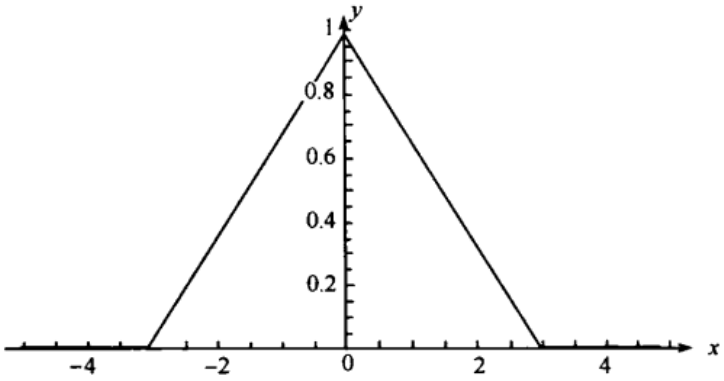


图 7.45 对称三角形核函数图示

其中，u 为可调参数，其默认值是 3，且 u 只取正实数。

图 7.46 给出柯西核函数 $y=\frac{1}{1+3x^2}$ 的图像。u 的大小决定了图像的“胖瘦”，u 越大，图像越“尖”，拟合误差越小，但容易造成“过学习现象”，影响推广能力。

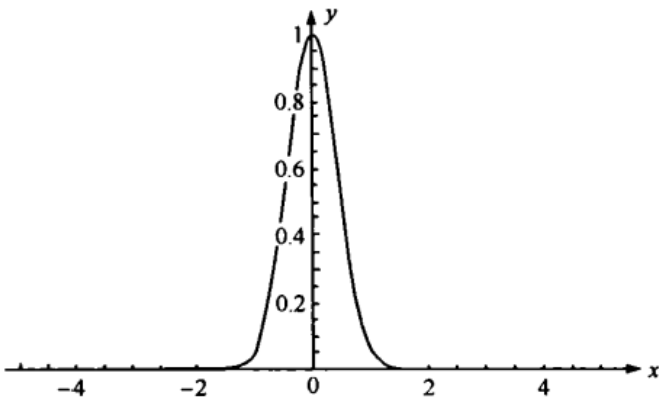


图 7.46 柯西核函数图示

7.7.5 拉普拉斯核函数(Laplace Function Kernel)

$$K(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n} \mathrm{e}^{-u|x_i-y_i|} \tag{7.11}$$

其中，u 为可调参数，其默认值是 3，且 u 只取正实数。

图 7.47 给出拉普拉斯核函数 $y=\exp(-3|x|)$ 的图像。u 的大小决定了图像的“胖瘦”，u 越大，图像越“尖”，拟合误差越小，但容易造成“过学习现象”，影响推广能力。

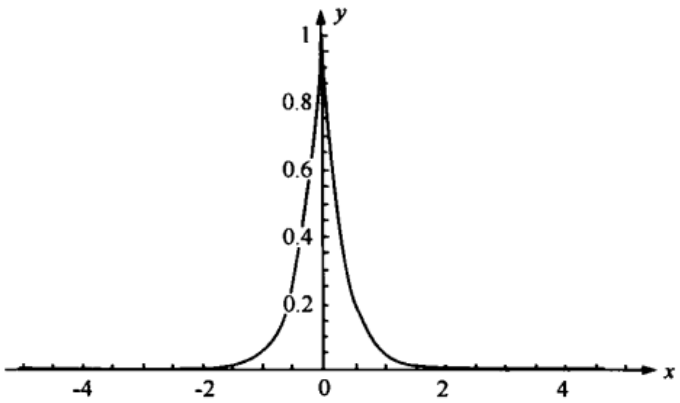


图 7.47　拉普拉斯核函数图示

7.7.6　双曲正割核函数(Hyperbolic Secant Function Kernel)

$$K(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{2}{e^{u(x_i-y_i)} + e^{-u(x_i-y_i)}} \tag{7.12}$$

其中，u 为可调参数，其默认值是 3，且 u 只取正实数。

图 7.48 给出双曲正割核函数 $y=\dfrac{2}{e^{3x}+e^{-3x}}$ 的图像。u 的大小决定了图像的“胖瘦”，u 越大，图像越“尖”，拟合误差越小，但容易造成“过学习现象”，影响推广能力。

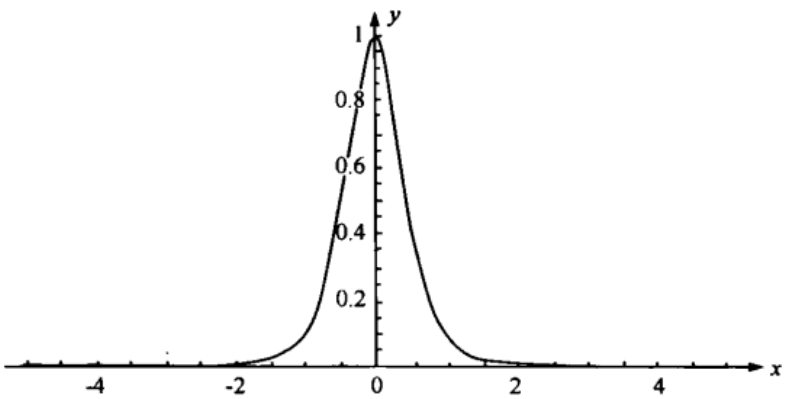


图 7.48　双曲正割核函数图示

7.7.7　平方正弦核函数(Squared Sin Cardinal Function Kernel)

$$K(\boldsymbol{X},\boldsymbol{Y}) = \prod_{i=1}^{n} \frac{\sin^2[u(x_i-y_i)]}{u^2(x_i-y_i)} \tag{7.13}$$

其中，u 为可调参数，其默认值是 3，且 u 只取正实数。

图 7.49 给出平方正弦核函数 $y=\dfrac{\sin^2(3x)}{(3x)^2}$ 的图像。u 的大小决定了图像的“胖瘦”，u 越大，图像越“尖”，拟合误差越小，但容易造成“过学习现象”，影响推广能力。

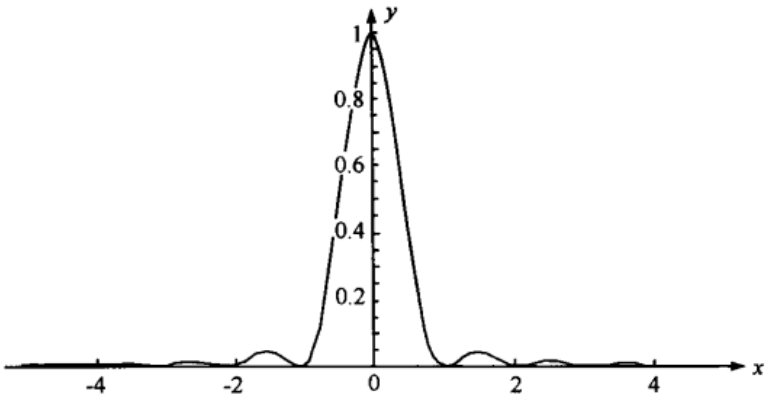


图 7.49 平方正弦核函数图示

图 7.44—7.49 中有几个粗看很相似，但它们代表了实质上不同的变化过程。CMSVM 系统允许用户补充自己新定义的核函数，当然定义要合理。

7.8 关于 CMSVM 系统中的其他参数

除了核函数及核参数外，应用 CMSVM 2.0 系统时还需要考虑确定如下的算法参数和控制参数。下面分别给出这些参数的含义，其在系统中的缺省值规定和取值的建议。

7.8.1 参数 *c*

在做“两类分类”和“函数估计”(包括交叉验证)时，总要设定参数 C 并要对其进行寻优。在 CMSVM2.0 系统中，C 的缺省值为 100。

在两类分类问题中，绝大多数情况下都会出现错分的样本。参数 C 是在寻找最优划分超平面计算时对被错分样本的一个惩罚系数。C 越大惩罚的力度越大，使训练集中被错分的样本越少，同时模型越复杂。在系统计算中，错分的样本也作为支持向量，且其在模型中对应的 α 系数等于 C，造成其也落在边界上的假象。

在函数估计问题中，总会出现落在“ε-带”以外的样本。参数 C 是在寻找最优回归超平面计算时，对落在“ε-带”以外样本的一个惩罚系数。C 越大惩罚的力度越大，使训练集中落在“ε-带”以外的样本越少，同时模型越复杂。在系统计算中，落在“ε-带”以外的样本也作为支持向量，且其在模型中对应的 α 系数等于 C，造成其也落在边界上的假象。

关于参数 C 值的选取，没有一般的规则可循。在函数估计问题中建议选取样本的输出值的取值范围的“宽度”作为 C 的初值。

7.8.2 函数估计中的控制参数

函数估计问题是比两类分类问题更为复杂的问题。原则上说，如果能够解决好函数估计问题，则两类以至于多类分类问题自然就解决了。因此函数估计问题的计

算量远远大于分类问题。为了提高计算效率，我们引入了如下几个参数。

(1) 回归带宽参数

回归带宽参数对应于寻求最优回归超平面时的带宽参数 ε，它的大小不但影响计算时间，而且对所建模型的推广能力有很大影响。一般来说，带宽参数越大，计算时间越短，回归函数越“平坦”，落在“ε-带”以外的样本越多。反之，带宽参数越小，计算时间越长，回归函数越复杂，落在“ε-带”以外的样本越少，容易造成“过学习”。

回归带宽参数的选取要根据实际问题的允许误差范围作具体分析。对于允许误差较大的实际问题，没有必要选择比允许误差更小的回归带宽参数。比如，预报温度时，若允许偏差可以在±1 ℃之间，则没有必要选取比 1 小的回归带宽。

在 CMSVM 2.0 系统中回归带宽的缺省值为 0.1。

必须指出的是，选取过小的回归带宽参数可能造成循环迭代计算次数过多，以至于得不出最优模型来。因此建议初始建立学习模型时不妨设置较大的回归带宽，将回归带宽参数的缺省值扩大 100～1 000 倍。

(2) 损失函数叠加上界

SVM 方法在回归建模时，采用了一种新的损失函数——ε-不敏感损失函数

$$L(y,f(x,a)) = L(|\ y-f(x,a)\ |_\varepsilon) = \begin{cases} 0 & \text{若}\ |\ y-f(x,a)\ | \leqslant \varepsilon \\ |\ y-f(x,a)\ |-\varepsilon & \text{其他} \end{cases}$$

则优化计算时要不断累加这一误差。显然，如果累加误差已经很大了，则当次计算的模型不可能为最优模型，因而没有必要继续进行计算。为了提高运算效率，我们对 ε-不敏感损失函数的累加误差限制了上界，这是因为当 ε-不敏感损失函数值过大时，所进行的函数估计值误差太大，这种运算结果是没有实际意义的，应立即停止其后的运算，进入重新开始的循环。

在 CMSVM 2.0 系统中损失函数叠加上界的缺省值为 2 000。

需要指出的是，因为函数不同和训练样本的差异，对 ε-不敏感损失函数值的大小界定也会不同，在处理实际问题时，需要根据实际问题进行多次试验，然后再设定合适的边界。

必须指出的是，选取过小的损失函数叠加上界参数可能造成程序过早终止，以至于得不出最优模型来。因此建议初始建立学习模型时不妨设置较大的损失函数叠加上界参数，比如 200 000。

(3) 回归迭代最大次数

CMSVM 系统建模程序在回归建模时，由于受计算机计算精度的影响，在处理大量数据时，经常出现迭代优化时间过长，无法确定支持向量的情况。为了提高解决实际问题的效率，我们采用了限制迭代次数的方法来提高工作效率，根据不同的问题限制了不同的迭代次数，凡是迭代次数超过规定的，就不再进行迭代计算，只把当前

的计算结果当作最终的迭代结果。

在 CMSVM 2.0 系统中回归迭代最大次数的缺省值为 10 000。而实际应用中要根据实际问题进行多次试验，设定合适的迭代次数。

必须指出的是，选取过小的回归迭代最大次数参数可能造成程序过早终止，以至于得不出最优模型来。因此建议初始建立学习模型时不妨设置较大的回归迭代最大次数参数，比如 1 000 000。

7.8.3 贴近度计算中的转换参数

在贴近度计算中主要是考察正、负样本间的贴近情况，对于分类问题这是很自然的。但是对于函数估计(回归)问题，本来只有一类样本点。为了对函数估计问题做样本数据分析，我们在回归问题的输出值域上引入 4 个转换参数 m,n,o 和 p(如图 7.50)。



图 7.50 转换参数图示

将输出函数值落在$[n,m]$中的样本点看作正样本，输出函数值落在$[p,o]$中的样本点看作负样本；也可以将输出值落在$[0,u]$中的样本点看作正样本，输出函数值落在$[p,0]$及$[n,m]$中的样本点看作负样本。这样就从所有的样本中转换出正、负两类样本。改变参数的值，就可以做不同的贴近度分析。

7.9 关于 CMSVM 系统中的参数寻优

在 SVM 方法中，参数的选取仍然是一个尚未解决的问题。我们在 CMSVM 2.0 中尝试提出了一种循环寻优的实用方法。

7.9.1 参数寻优的基本思路

对于固定的训练数据，给定一组参数就可以应用 SVM 方法建立一个学习机，随着参数的改变学习机也跟着改变，并且所得到的学习机都是在特定条件下的全局最优解。

在 CMSVM 2.0 的应用中，由于参数(们)的循环改变，可能建立了成百上千个学习机。如何在众多的学习机模型中选出针对实际问题而言最有效的一个，这就是参数寻优的目的。

建模和寻优都必须(只能)依据已有的数据，因此必须将全部数据切割为训练数据和寻优数据——用一部分数据来建模，用其余的数据来优化参数。

在“两类分类”和“函数估计”选项中是将全部数据样本随机分割为三个部分：用于建模的训练样本子集(约占样本数的 75%)、用于优化模型参数的实验样本子集

(约占样本数的 20%)和用于检验模型推广能力的检验样本子集(约占样本数的 5%)。其中训练样本子集用于直接训练学习建模,所建的每一模型均对实验样本子集做预测。按照指定的模型择优标准选出最优者,作为最优模型,再应用此模型对检验样本子集做预测。由于检验样本子集中的样本从未参与建模和选优,它独立于建模和参数优化过程,所以用它得到的预测结果比较有代表性。

"交叉验证 1"和"交叉验证 2"选项是在数据集中随机抽取一部分数据作为训练文件建模,再用另一部分数据做效果验证。这个过程通过对训练样本的多次随机抽取而交叉进行。显然,它相当于进行多次"两类分类"和"函数估计"建模。进行的次数越多,所得到的结果就越具有代表性。

在"交叉验证 1"中,每次抽取一个样本作为待验证的检验样本,用剩余的所有样本建模,顺序对所有样本循环此过程,适用于只有几十个样本的学习问题。

在"交叉验证 2"中是按任意次数、任意比例(百分数)随机把样本数据集切割成训练样本和验证样本,使验证更具代表性,它适用于一般的学习问题。

假定应用于同样的训练数据,由不同参数建立了两个学习机。如何在两者之中择优?择优要依据一定的标准,这就产生一个挑选学习机的择优依据或择优标准的问题。"两类分类"和"函数估计"采用的是不同的择优标准。

参数寻优就是依据给定的择优标准,在所建立的所有备选学习机中选出一个最优者(注意:如果依据给定的择优标准得出的最优者不唯一,则需要进一步引入择优标准)。

7.9.2　分类问题的择优标准

CMSVM 2.0 给出 3 种两类分类的模型择优标准(如图 7.12 所示),它们是:正样本 TS 评分、负样本 TS 评分和全体样本准确率。

当两类样本的发生频率接近(即正、负训练样本数几乎相同)时,可以选取全体样本准确率进行择优,算法是

$$\text{全体样本分类预报准确率} = \frac{\text{分类正确的样本数}}{\text{所有样本数}} \times 100\%$$

当正类(负类)样本的发生频率很小时,应该选取正(负)样本的 TS 评分进行择优。正、负样本的 TS 评分的算法是

$$\text{正样本 TS 评分} = \frac{\text{预报为正类且确实是正类的样本数}}{\text{正类中所有的样本数} + \text{负类中被预报为正类的所有样本数}} \times 100\%$$

$$\text{负样本 TS 评分} = \frac{\text{预报为负类且确实是负类的样本数}}{\text{负类中所有的样本数} + \text{正类中被预报为负类的所有样本数}} \times 100\%$$

在分类运算中,根据应用实际,我们采用了两级 TS 评分机制,以期选择更稳定的模型。即首先选用实验文件 TS 评分标准当作评选最优模型的标准,选择使实验文件 TS 评分值达到最大的模型为最优模型;当出现多个模型使实验文件 TS 评分达

到最大值时，再选用训练文件的 TS 评分当作评选最优模型的第二级标准，选取这些模型中使训练文件的 TS 评分值最大的模型为最优模型。

7.9.3 关于分类精度中几个名词的解释

评判一个分类模型的好坏，除了在 7.9.2 中提到的正样本 TS 评分、负样本 TS 评分和全体样本准确率外，在分类问题的最优模型和计算结果文件的查询中还出现了“预报成功率 (Precision)”和“预报概括率 (Recall)”两个名词。

当以正样本的 TS 评分作为寻优标准时，它们的计算公式是

$$\text{正样本的预报成功率} = \frac{\text{预报为正类且确实是正类的样本数}}{\text{预报为正类的所有样本数}} \times 100\%$$

$$\text{正样本的预报概括率} = \frac{\text{预报为正类且确实是正类的样本数}}{\text{正类中所有的样本数}} \times 100\%$$

当以负样本的 TS 评分作为寻优标准时计算公式类似。通过图 7.51 对上述各种分类精度定义加以直观解释。

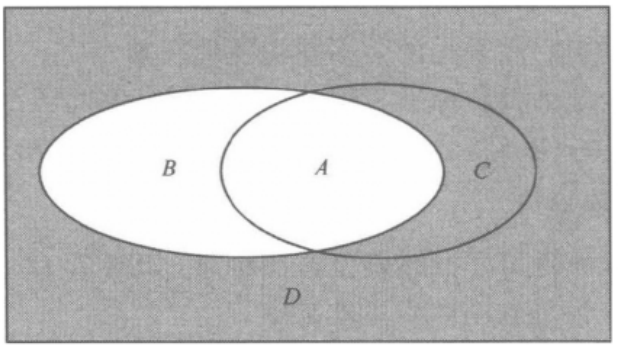


图 7.51 分类精度定义图示

任一两类分类预报的实际问题都可以通过图 7.51 表示，其中的矩形框表示全部样本，它被划分成四个子区域 A，B，C 和 D。其中白色椭圆区域 $(A+B)$ 为实际的正类样本，阴影区域 $(C+D)$ 为实际的负类样本。另一个椭圆区域 $(A+C)$ 为预报为正样本的区域，它的余集 $(B+D)$ 是预报为负样本的区域。假定我们用 $|A|$，$|B|$，$|C|$ 和 $|D|$ 分别表示对应区域所含样本的个数，则有：

$|A|$ = 实况是正样本且预报是正样本的个数；

$|B|$ = 实况是正样本且预报是负样本的个数；

$|C|$ = 实况是负样本且预报是正样本的个数；

$|D|$ = 实况是负样本且预报是负样本的个数。

这样我们可以得到

$$\text{全体样本分类预报准确率} = \frac{|A|+|D|}{|A|+|B|+|C|+|D|} \times 100\%$$

$$\text{正样本 TS 评分} = \frac{|A|}{|A|+|B|+|C|} \times 100\%$$

$$\text{负样本 TS 评分} = \frac{|D|}{|B|+|C|+|D|} \times 100\%$$

$$\text{正样本的预报成功率} = \frac{|A|}{|A|+|C|} \times 100\%$$

$$\text{正样本的预报概括率} = \frac{|A|}{|A|+|B|} \times 100\%$$

$$\text{负样本的预报成功率} = \frac{|D|}{|B|+|D|} \times 100\%$$

$$\text{负样本的预报概括率} = \frac{|D|}{|C|+|D|} \times 100\%$$

可以看出正、负样本的 TS 评分是较为严格的标准，它兼顾了错预报和漏预报两个方面。

7.9.4　回归模型的择优标准

CMSVM 2.0 给出四种回归模型择优标准(如图 7.23 所示)，它们是：绝对差、相关系数、均方差和距平。

(1)绝对差：是计算预测值与实际值之差的绝对值的平均值，计算公式为

$$\text{绝对差} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \tilde{y}_i| \qquad (\tilde{y} \text{ 是对 } y_i \text{ 的预测值})。$$

(2)均方差：是计算预测值与实际值之差的平方的平均值的开方值，计算公式为

$$\text{均方差} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n}} \qquad (\tilde{y} \text{ 是对 } y_i \text{ 的预测值})。$$

(3)相关系数：是计算预测值与实际值的相关情况，计算公式为

$$\text{相关系数} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(\tilde{y}_i - \overline{\tilde{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(\tilde{y}_i - \overline{\tilde{y}})^2}},$$

其中，$\tilde{y}_i$ 是对 y_i 的预测值，$\overline{y}$ 是 y_i 的平均值，$\overline{\tilde{y}}$ 是 $\tilde{y}_i$ 的平均值。

(4)距平：是一种粗略的估计值，它不管预测值与实际值的具体差值大小，而只关心与实际均值的偏离趋势——计算预测值与实际值关于实际均值的偏离趋势相同的样本占总样本的比例，计算公式为

$$\text{距平} = \frac{(y_i - \bar{y}) \text{ 与 } (\tilde{y}_i - \bar{y}) \text{ 符号相同的样本数}}{\text{总样本数}},$$

其中，$\tilde{y}_i$ 是对 y_i 的预测值，$\bar{y}$ 是 y_i 的平均值。

用户可以根据实际问题的需要，在上述四种评分标准中任意选择一种作为评选最优回归模型的标准。

7.9.5 交叉验证的择优标准

交叉验证只是重复“两类分类”和“函数估计”的计算过程，所以采用的是 7.9.2 和 7.9.4 中给出的方法。

7.10 本章小结

(1) 简单介绍了中文计算机学习建模软件平台 CMSVM 2.0 的设计思路及建模参数寻优方法。

(2) 介绍了 CMSVM 2.0 的基本功能，安装所需的软、硬件环境及过程。

(3) 介绍了 CMSVM 2.0 所需的数据文件格式。

(4) 介绍了 CMSVM 2.0 提供的“两类分类”、“函数估计”、“预报应用”、“贴近度计算”、“交叉验证 1”、“交叉验证 2”及“文件操作”等功能的使用方法，以及上述功能运行生成的各类文件及其查询方法。

(5) 简单介绍了 CMSVM 2.0 系统中给出的 7 种核函数及其使用参数。

(6) 简单介绍了 CMSVM 2.0 系统的寻优参数、控制参数、转换参数等的定义及使用方法。

第8章　支持向量机方法在天气预报中的应用

近十几年来，随着大气科学本身及计算机等相关科学的迅速发展，天气预报技术有了长足的进步。天气预报技术的发展史就是一部以适应社会发展为动力，以其他领域科学技术的进步与相互渗透为条件，以预报手段日益进步、预报准确率不断提高为标志的历史。概率论和统计学的进展及其在气象领域中的应用，促使继天气学、动力学方法之后现代天气预报的第三种预报方法——统计天气预报方法的产生和发展。多层递阶预报(Multi-level Recurtive Prediction Method，MLRPM)、灰色预测(Gray Prediction，GP)、动态相似统计预报(Dynamical Analogue Statistical Prediction，DASP)等基于统计学习理论的具体预报技术相继开发出来，完全预报(Perfect Prognosis，PP)、模式输出统计预报(Model Output Statistical，MOS)和卡耳曼滤波(Kalman Filtering，KF)等技术更成为数值预报产品定量应用制作气象要素的主要方法。此外，人工智能和自动化技术的进展也对现代天气预报技术的发展起着至关重要的作用。气象预报专家系统、人工神经网络等基于人工智能技术的新型预报工具的相继问世，使传统的经验预报迈入现代化的行列。

常规统计方法建立的基础是基于大数定律的因子与预报对象之间的显著线性相关，建模都有一个特点：建模之前，必须事先确定或假设模式里的因变量和自变量之间的函数关系 f，实际上由于大气变化的复杂性和非线性性，使得因变量与自变量之间的关系呈现复杂的非线性关系，所以这种事先给定的函数关系不可能较好地描述因变量与变量之间的联系。

任何天气现象或气象要素的演变，实质上都是前期某些气象要素有条件的组合作用影响的结果，而这种因子组合又是形式多样、繁杂的非线性组合，即便一个因子值固定，只要其他因子的值产生各种变化，也会表现出多种多样的结果，而各个气象要素的变化彼此不一致，是随机的，因此要用因子之间的函数构造来表述预报对象，显然就十分困难。而如何来表述这种由多种因素构成的空间、甚至是时间的特定组合环境下，产生的天气或气象要素的变化，就是我们气象预报工作者致力于解决的问题。SVM 方法的出现为我们提供了一个解决预报问题的新方法。

SVM 建模方法，其本质就是通过对各种典型空间(支持向量)的充分表述来描述因子群与预报对象之间的关系，而不是通过单一因子的系数变化来刻画预报对象，这与基于因子权重的常规统计预报方法有明显的不同，近似于专家系统的规则推理。然而，专家系统的各种规则是基于专家的知识来表达的，依赖于专家的知识积淀，而 SVM 预报模型却是通过自学习的方式获得，人为的先验干预较少，是一种基于事实的转导式推理。本章主要介绍 SVM 在短期天气预报和短期气候预测领域应用的情况。

8.1 SVM 方法用于降水预报和温度预报的建模实例

在我国天气预报中，SVM 最早应用于降水和温度预报(冯汉中等 2004)，下面简要介绍这一工作。

8.1.1 SVM 推理预报模型的建立方法

虽然 SVM 方法同基于概率测度和大数定律的传统统计方法风格迥异，但它与通常的统计分析预测方法有着天然的联系(有大致相同的问题表述和数据预处理方式)，预报模型仍然是建立在大量样本资料的基础之上。建模步骤与常规方法基本相同：

(1) 创建样本资料；

(2) 构造预报因子；

(3) 因子的归一化处理；

(4) 确定建立 SVM 模型所使用的核函数；

(5) 创建训练集；

(6) 创建试验集；

(7) 确定寻求最优 SVM 推理预报模型的评定准则；

(8) 通过评定准则确定最终用于预报的 SVM 推理模型。

8.1.2 在降水分类预报和温度回归预报试验中 SVM 的应用

8.1.2.1 样本资料和预报对象

随着大气环流的不断变化，大气的空间状态在不断改变，人们所关注的预报对象也在随之产生变化，因此，要较好地对预报对象进行估计，就需要比较充分地描绘大气的空间状态及其随时间的变化特征。由于 SVM 的最终决策函数只由支持向量确定，复杂程度取决于支持向量的数目，而不是样本空间的维数(因子数)，故人们可以选取较多的因子以充分表述大气的时空变化。然而，尽管 SVM 方法是基于小样本理论建立的，但样本太少不可能反映出气象要素的气候变化规律，因此，在选取预报因子来描述大气状态的时空变化时，还要准备足够多的样本资料，因而选用有较多样本资料的欧洲中期天气预报中心(European Centre for Medium-Range Weather Forecasts，ECMWF)的输出产品，利用其 500 hPa 高度、850 hPa 温度、地面气压的 0 小时资料来构造预报因子。针对影响四川盆地天气变化的关键区域，在不同的层次选取不同的区域来构造预报因子(如图 8.1)。因子的主要构成方式为：所选区域的值、24 小时变化量、相关区域之间的差值等，共构成预报因子 77 个，尽量以数字化形式来描述系统及要素的演变，避免人为判断因素的参与。为了避免各个因子之间的量差异，对全部样本的每一个因子分别做归一化处理，使每一个因子的数据均落入区间[0,1]。

(1) SVM 分类预报试验对象——四川盆地片区面雨量

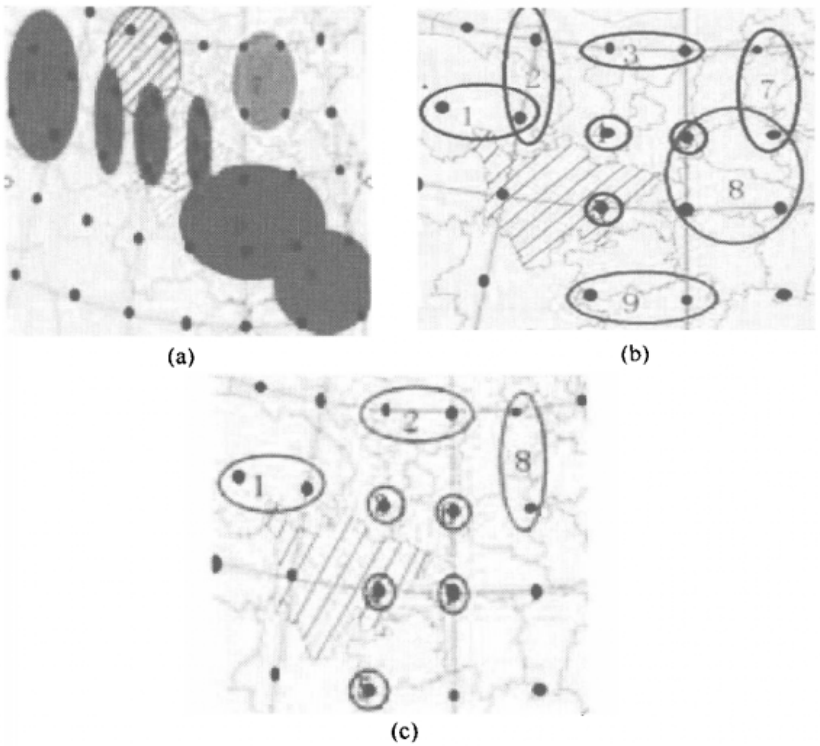


图 8.1　不同层次选取不同要素构造预报因子

(a)500 hPa 预报因子选取;(b)地面气压预报因子选取;(c)850 hPa 温度预报因子选取

如图 8.2 所示,考虑四川盆地降雨的气候特点,将四川省内盆地划分为三个片区:①盆地西北部、②盆地东北部和③盆地西南部,分别作为预报对象进行试验。这里关注的是大于 15 mm 降水的面雨量(所选区域内站点雨量的平均值),因此把面雨量大于或等于 15 mm 的归为 +1 类,小于 15 mm 的归为 −1 类。

(2) SVM 回归预报试验对象——四川盆地内单站日平均气温

由于气团具有相对稳定和均匀的性质,所以单站气温对邻近区域具有较好的代表性,在四川盆地内选取了 6 个站点:成都(56294)、绵阳(56196)、宜宾(56492)、达川(57328)、南充(57411)及内江(57504),分别作为预报对象进行试验。

资料长度为 1990—2000 年共 11 年(4—9 月)ECMWF 的 0 小时历史资料及对应的温度、雨量资料。并将其分为了三部分:训练集、试报集和检验集。建立检验集是为了对最终确立的 SVM 预报模型做预报,以检验 SVM 模型的预报效果(推广能力),检验集的数据不参与训练学习及参数筛选等建模过程。

8.1.2.2　建模方式

以径向基函数作为核函数建立推理试验模型,径向基函数形式参见式(7.8)。

尽管选定径向基函数作为建立 SVM 模型的核函数,但随着参数值选取的不同,函数形态就会发生相应的变化,进而引起 SVM 模型的变化。以分类为例,随着参数 g 和 b 的改变,迭代次数变化明显:少的几千次,多的达 10 万次以上;错分样本差异

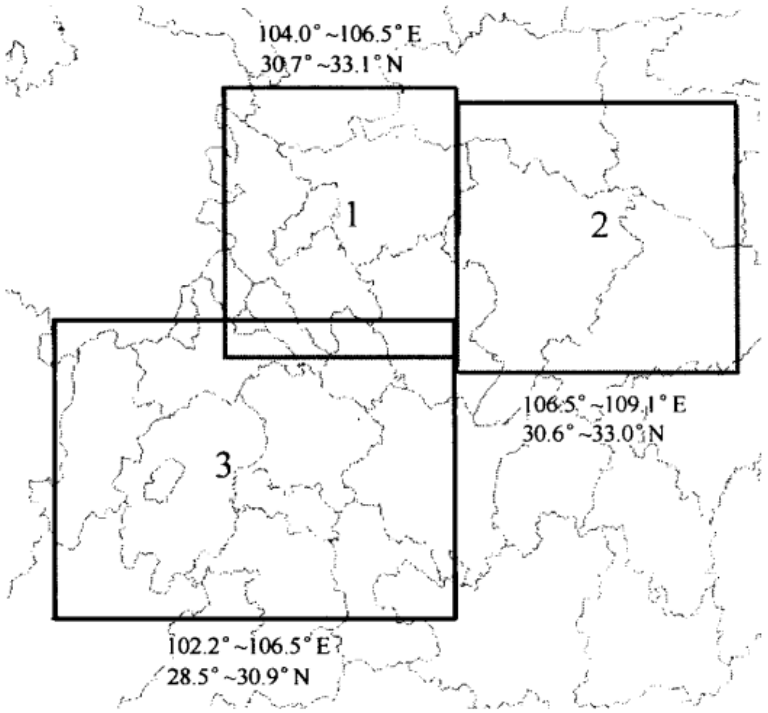


图 8.2 四川盆地面雨量区域划分图示

大:少则一个,多则上百;运行时间长短不一:十几秒到近 100 秒不等;支持向量、VC 维、核计算次数的变化也十分显著。因此,选取不同参数值建立的 SVM 模型,其结果有较大的差异,对应的预测结果(如 TS 评分)也大不一样。

采用逐步筛选的方法确定参数:首先设置较大的参数取值范围,对参数进行大间隔步长的循环取值,通过训练和测试,依据评分准则确定在这种状况下建立的最优 SVM 预报模型所对应的参数值,再以此参数值为中心,设置较小的参数范围,以小间隔步长重复上述步骤,直至最终确定出用于建立 SVM 预报模型的参数值,进而确定预报模型。径向基核参数 g 和 b,以及支持向量系数等都可由 CMSVM 系统自动生成,不需人为选取。

8.1.2.3 SVM 模型的试验效果分析

(1) 分类试验效果分析

在分类试验中,建立样本集的方式为:在 1990—1998 年 4—9 月的样本中随机剪取了近 100 个样本作为检验集,把其中剩余的近 1 400 个样本作为训练集,而将 1999—2000 年 4—9 月接近 350 个样本作为测试集。

表 8.1 为通过参数逐步筛选训练、而在试报集中获得最佳 TS 评分的 SVM 预报模型对训练集、试报集、检验集进行预报的评分结果。

表 8.1　四川盆地面雨量(大于 15 mm)SVM 预报模型试验结果

四川盆地分片区域	检验项目结果显示	有雨样本数(个)	样本总数(个)	频率	正确数目(个)	空报数目(个)	漏报数目(个)	TS 评分	m 正确数目(个)	m 空报数目(个)	m 漏报数目(个)	m 评分
盆地西北部，支持向量为 231；$g=0.031\ 25$，$C=1\ 251$；	对训练集的回报	99	1 456	0.068	84	104	15	0.413 8	114	74	15	0.561 6
	对试验集的预报	24	336	0.071 4	7	31	17	0.127 3	13	25	17	0.236 4
其中有雨的支持向量为 96	对检验集的预报	10	92	0.108 7	5	5	5	0.333 3	7	3	5	0.466 7
盆地西南部，支持向量为 352；$g=0.625$，$C=101$；	对训练集的回报	113	1 439	0.078 5	105	13	8	0.833 3	110	8	8	0.873 0
	对试验集的预报	25	336	0.074 4	15	37	10	0.241 9	28	24	10	0.451 6
其中有雨的支持向量为 105；	对检验集的预报	10	109	0.091 7	6	12	4	0.272 7	10	8	4	0.454 5
盆地东北部，支持向量为 345；$g=0.81$，$C=151$；	对训练集的回报	131	1 452	0.090 2	125	7	6	0.905 8	130	2	6	0.942
	对试验集的预报	31	336	0.092 3	14	42	17	0.191 8	23	33	17	0.315 1
其中有雨的支持向量为 114	对检验集的预报	13	96	0.135 4	7	12	6	0.280 0	13	6	6	0.52

注：表中带 m 部分为以预报出现的当天或前后一天有实况出现则计为正确的预报评定

从表 8.1 中可以看出：

盆地西北部：训练集样本为 1 456 个，试报集样本为 336 个，检验集样本为 92 个。通过训练集和试报集的参数逼近，最终获得推理模型的支持向量为 231 个，占总样本的 16%，其中有雨样本 96 个，占有雨分类样本(99 个)的 97%。

盆地西南部：训练集样本为 1 439 个，试报集样本为 336 个，检验集样本为 109 个。通过训练集和试报集的参数逼近，最终获得推理模型的支持向量为 352 个，占总样本的 24%，其中有雨样本 105 个，占有雨分类样本(113 个)的 93%。

盆地东北部：训练集样本为 1 452 个，试报集样本为 336 个，检验集样本为 96 个。通过训练集和试报集的参数逼近，最终获得推理模型的支持向量为 345 个，占总样本的 24%，其中有雨样本 114 个，占有雨分类样本(131 个)的 87%。

以上分析说明：用 SVM 方法对三个片区的面雨量建立的推理模型，其支持向量仅占样本数的 20%左右，即对样本进行了高度的浓缩，通过训练去除了冗余的样本集，而支持向量中又较好地包含了所关注的面雨量事件。对盆地东北部和西南部训练集的回报 TS 评分高达 0.8 以上，对试报集预报的 TS 评分为 0.2 左右，而对检验集预报的 TS 评分达到了 0.27 以上，高于试报集的 TS 评分，且均高于有雨样本出现的频率，具有正预报技巧，说明对应的模型具有较好地预报推理能力。而对盆地西北

部训练集的回报表明，尽管漏报少，但空报较多，导致 TS 评分较低，间接说明选取的预报因子难以区分关注的预报对象，应对预报因子进行挑选或另外构造预报因子，尽管如此，其 TS 评分仍高于有雨样本出现的频率，具有正的预报技巧。

如果以预报出现的当天或前后一天有实况出现计为正确的话，尽管漏报没有什么变化，但空报减少了，TS 评分明显提高（表 8.1 中带 m 的部分），说明推理模型对预报对象的演变有较好地把握能力。

（2）回归试验效果分析

在回归试验中建立样本集的方式为：依据计算机程序，设置临界值，通过随机数发生器进行样本的随机采集，将 1990—2000 年 4—9 月间资料中的 1 900 个样本分为三部分：训练集（约 1 250 个样本）、试验集（约 350 个样本），以及检验集（约 300 个样本）。

表 8.2 为通过参数逐步逼近训练、在试验集中获得最佳相关系数的预报模型进行预报的结果评价。包括三部分：对训练集的回报检验、对试验集的预报检验、对检验集的预报检验。

表 8.2 用 SVM 回归方法作四川盆地内单站逐日平均气温预报试验结果（g,C,w 为建模参数）

试验站点	试验项目	样本数（个）	SV 数（个）	相关系数	均方根误差（℃）	绝对误差（℃）
绵阳	对训练集的回报	1 217		0.963 0	1.079 8	0.839 5
g=0.077，C=190	对试验集的预报	382	953	0.946 8	1.237 1	0.956 2
w=0.343	对检验集的预报	285		0.944 0	1.297 4	0.999 2
成都	对训练集的回报	1 254		0.959 2	1.091 2	0.836 8
g=0.046，C=300	对试验集的预报	364	1 042	0.953 2	1.137 2	0.886 6
w=0.257	对检验集的预报	266		0.922 7	1.382 6	0.978 9
宜宾	对训练集的回报	1 212		0.964 3	1.051 6	0.887 8
g=0.28，C=100	对试验集的预报	373	285	0.939 5	1.351 7	1.046 3
w=1.48	对检验集的预报	299		0.957 7	1.155 8	0.905 1
达川	对训练集的回报	1 230		0.968 3	1.160 7	0.863 8
g=0.163，C=50	对试验集的预报	361	1 177	0.953 5	1.303 9	1.043 0
w=0.078	对检验集的预报	293		0.949 6	1.318 8	1.023 4
南充	对训练集的回报	1 212		0.973 5	1.019 4	0.776 6
g=0.078，C=470	对试验集的预报	390	961	0.962 0	1.242 1	0.953 9
w=0.343	对检验集的预报	282		0.964 5	1.197 3	0.939 3
内江	对训练集的回报	1 255		0.962 6	1.148 5	0.884 4
g=0.155，C=50	对试验集的预报	381	996	0.950 3	1.285 9	1.012 8
w=0.327	对检验集的预报	248		0.958 4	1.317 8	1.042 9

尽管采用的是相同的预报因子群来对应不同站点的逐日平均气温进行 SVM 预报模型的建立，但建立的各个站点的预报模型的支持向量却有差别，建立预报模型的参数也有明显的差异。宜宾站支持向量仅为 285 个，大大少于建模样本数，其余站点的支持向量位于 950～1 200 个，也就是说为了达到一定的预报准确度，支持向量回

归方法是通过一定数量的支持向量来描绘预报对象的非线性演变过程。从表 8.2 中可以看出，在每一个站点中，均是对训练集的回报效果最好，相关系数保持在 0.95 以上，绝对误差在 0.9 ℃以内。对试验集和检验集的预报评价表明了 SVM 回归预报模型的稳定性，对所有站点的检验集的预报检验显示，相关系数均在 0.92 以上，绝对误差保持在 1.1 ℃以内，均方根误差也小于 1.4 ℃。

8.1.3　试验结果小结

应用 SVM 方法建立降水和温度预报的试验结果表明：

(1) SVM 是通过对支持向量样本“加权”的方式构造推理模型，而不是对每一个因子进行加权，因而对因子的数量没有明显的限制。因此，通过选取与预报对象有明确意义的各种因子，就可以较充分地表述预报对象与预报因子之间变化的时间和空间概念，这一特点明显优于传统的统计预报方法。

(2) SVM 方法是通过对历史样本的不断学习来建立预报模型的，SVM 预报模型的建立过程就是对历史样本的自学习记忆过程。试验表明，尽可能多的样本参与训练，能使预报模型包含比较完备的支持信息，从而使得预报更准确。因此，在实际预报中要不断增加新的样本参与训练，进而达到不断改善预报模型的目的。

8.2　SVM 方法用于预报因子筛选的实例

8.2.1　交叉验证优化参数

由于 SVM 是基于历史数据训练学习的一种建模方法，当其使用不同的训练样本、不同核参数时，得到的预报模型是不同的。如何筛选预报因子、优化参数使得预报模型最优，一直是 SVM 方法应用的难点，也是该方法使用者关注的问题。这里以天空云量为预报对象，基于 SVM 方法进行大量的交叉验证来筛选预报因子和优化核参数。

8.2.1.1　训练样本集的构建

天空云量是云遮蔽天空视野的成数，即有云天空占天空总面积的比例。我国云量计量单位采用 10 成制。天空无云，云量为 0；天空完全为云所遮蔽，云量为 10。以武汉市(台站号：57494)单站日平均总云量(每日 02，08，14 和 20 时 4 个时次总云量的平均值)为预报对象，约定日平均总云量＜4 为“晴天到少云”。由于武汉市位于亚热带季风区，不仅受中高纬西风带系统的影响，而且还受西南季风、东南季风等系统的影响，选取预报因子时既要尽可能地描述大气的运动和变化状态，又要考虑可以得到足够的样本长度。所以试验选取了 2001 年 5 月 1 日—2004 年 12 月 31 日(共 1 340个样本)逐日武汉市地面 02，08，14，20 时的气温、相对湿度、气压、风、总云量、低云量观测值和高空 08，20 时 925，850，700，500，400 hPa 的位势高度、温度、露点、风的观测值，以及 20 时 ECMWF 500 hPa 高度、850 hPa 温度、地面气压 24 小时预报场

及其组合场等(共 127 因子)历史资料作为训练学习分析数据。

由于待选的预报因子太多,为了减少筛选因子的盲目性,因而对上述资料进行了处理,即在训练样本中分别去掉一些因子来重新构造样本,形成不同的训练样本集。现只介绍下述 5 类:

(1)训练样本集 1:包含了上述所选的资料,共有 127 个因子;

(2)训练样本集 2:去掉训练样本集 1 中 ECMWF 的组合场因子,保留其他因子,共有 114 个因子;

(3)训练样本集 3:由 ECMWF 24 小时预报场及其组合场和地面、高空资料中与湿度有关的量组成,共有 63 个因子;

(4)训练样本集 4:由 ECMWF 24 小时预报场及其组合场和 20 时的地面、高空资料组成,共有 81 个因子;

(5)训练样本集 5:由 ECMWF 24 小时预报场及其组合场组成,共有 49 个因子。

这样就形成了包含因子个数不同的 5 类训练样本集,需要说明的是,5 类训练样本集中都加入了后一天武汉市日平均总云量的实况值。

8.2.1.2 预报因子的筛选和参数的优化

应用 SVM 的两类分类方法,标定日平均总云量<4(晴到少云,占总样本的 21.27%)为正样本(+1 类)、日平均总云量≥4 为负样本(−1 类);从 5 类训练样本集的 1 340 个样本中各随机抽取 10%作为检验样本,其余 90%样本作为建模样本,来分别对建立的"晴天到少云"的预报模型进行交叉验证,为了避免各预报因子之间量级的差异,在建模之前,对全部样本的每个因子分别做了归一化处理,使每个因子的数据在[0,1]之间。

这里选取最常用的径向基核函数,通过调整核参数 g 的值(惩罚系数 $C=100$),进行大量随机交叉验证,分析比较所建模型 TS 评分的高低,从而选择出包含最佳预报因子的训练样本集。

表 8.3 仅给出了 g 为 0.001,0.01,0.032 和 0.05 时,按正样本的 TS 评分寻优标准进行的各 50 次随机交叉验证的统计结果。

从表 8.3 的结果可以看出,训练样本集 3 和 5 在 g 的 4 种取值下平均的 TS 评分值都比较低,最不适合作为建模训练样本集;而 g 为 0.001 时,对于 5 类训练样本集平均的 TS 评分都很低,特别是用训练样本集 5 建立的模型完全没有预报能力,所以它不是最优的核参数;当 g 取 0.01,0.032 和 0.05 时,训练样本集 1,2,4 的平均 TS 评分相对较高,其中以 g 取 0.01 和 0.032 时,训练样本集 4 的平均 TS 评分最高,因此可以认为 g 取 0.01 和 0.032 时的训练样本集 4 中包含了最佳的预报因子,即可选择训练集 4 作为下一步预报建模的训练样本。

表 8.3　5 类训练样本集各进行 50 次交叉验证的 TS 评分平均值

训练样本集 \ g 的取值	0.001	0.01	0.032	0.05
训练样本集 1	0.356 5	0.414 2	0.416 3	0.417 0
训练样本集 2	0.325 2	0.417 6	0.396 7	0.402 1
训练样本集 3	0.132 0	0.348 0	0.365 6	0.362 5
训练样本集 4	0.338 2	0.420 5	0.421 8	0.417 1
训练样本集 5	0.000 0	0.256 1	0.314 7	0.295 6

此外，不仅可以从 TS 评分的高低来选择最佳预报因子和核参数，还可以从边界上支持向量的个数、VC 维等方面进行核参数和预报因子的寻优。一般说来，TS 评分越高、边界上支持向量个数越多、VC 维越小，得到的预报模型越好。如表 8.3 中训练样本集 1，2，4 在 g 取 0.05 时平均的 TS 评分不是太低，但它们与之对应的 VC 维较大、边界上支持向量个数相对较少，所以选择 g 取 0.01，0.032 时 TS 评分较高的训练样本集 4 比较合理。限于篇幅，这里不一一列出边界上支持向量和 VC 维等数据。

从天气意义上来分析，也可以解释经过大量交叉验证筛选后的训练样本集 4 中包含了最佳的预报因子，该训练样本集既含有离预报时效较近的地面、高空因子，又含有预报场及其变化量，这些因子较全面地反映了大气的运动状态和变化趋势。而训练样本集 1 包含了太多前期的无用因子，增加了杂噪；训练样本集 2 既包含了太多前期的无用因子，使得干扰因素多，又没有反映预报场变化的组合因子，因而不能充分描述大气运动的变化；训练样本集 3 的地面、高空资料中只含有与湿度有关的量而没有温度、风等要素，对大气的热力和动力状况描述不足；训练样本集 5 仅由 ECMWF 数值预报产品组成，不包含直接反映大气前期状态的因子，会对预报模型的推理效果产生影响。因此训练样本集 1，2，3 和 5 都不是最佳的，只有训练样本集 4 才是最佳的建模训练样本。

从上面的分析可知，训练样本集 4 为最佳的建模训练样本。为了进一步确定模型的稳定性和核参数 g 的取值，利用训练样本集 4 作为学习建模样本，选取表 8.3 中初步选出的核参数，从中随机抽取 10%，20%和 30%的样本作为检验样本，分别对日平均总云量小于 4 的预报模型进行多次交叉验证以检验模型的稳定性及确定核参数；然后将训练样本集 4 的全部样本作为建模训练样本，用 SVM 方法建立日平均总云量小于 4 的预报模型，来试报 2005 年 1 月 1 日—5 月 31 日逐日 20 时武汉市逐日的日平均总云量。交叉验证和试报结果表明（表略），对于日平均总云量小于 4 的模型，最优的参数为：$C=100$，$g=0.01$。

8.2.1.3　小结

通过建立不同的训练样本集，基于 SVM 方法进行大量多因子的随机交叉验证，

来实现预报因子的筛选和核参数 g 的优化，为筛选预报因子和优化参数提供了思路。

8.2.2 比较支持向量样本

8.2.2.1 构建预报对象和预报因子

在全国选取了哈尔滨、大连、北京、郑州、武汉、南京、合肥、杭州、南昌、福州、厦门、广州、南宁和海口等 14 个站 08—08 时的 24 小时降水量，考虑到南北降水的差异，将哈尔滨、大连、北京 3 站 24 小时降水量≥1 mm 作为正样本(有雨样本)、24 小时降水量<1 mm 作为负样本；而其他 11 站 24 小时降水量≥10 mm 作为正样本(有雨样本)、<10 mm 作为负样本来对预报对象进行分类。

将 T213 数值预报产品和经过计算后的物理量场共 888 个因子分别插值到上述 14 个站中，并初步选取了与降水有关的 79 个因子作为预报因子，它们是海平面气压、涡度及散度、假相当位温、U 偏差风和 V 偏差风、湿位涡倾斜发展判据、K 指数、沙氏指数、累积水汽通量和累积水汽通量散度、动力综合因子、相对湿度的垂直累积量、垂直速度累积量、螺旋度累积量、相对湿度垂直差、散度垂直差、厚度平流垂直差、累积平均垂直速度等。同时由于数值预报可能存在着系统性偏快或偏慢的现象，而且在不同地区常有不同的表现，所以采用了前后预报时效的数值产品同时使用的方法，即选取了起始场为 20 时、预报时效为 0，12，24，36 小时的 4 个时次的数值产品。这样每个样本就包含了 316 个预报因子。为了避免各预报因子之间量级的差异，对全部样本的每个因子分别做了归一化处理，使每个因子的数据在[0，1]之间。

以 2003—2005 年共 3 年 4—9 月逐日的降水量实况和相应的 T213 数值预报产品的历史资料作为训练学习分析数据(实际只有 489 天资料)。

8.2.2.2 SVM 参数优化和初步建模

这里也应用 SVM 的两类分类方法，选取最常用的径向基核函数，通过调整惩罚系数 C、核参数 g 的值，分别进行大量多因子的随机交叉验证。比较所建模型 TS 评分的高低，从而选择出最佳的参数 C 和 g，建立最优预报模型；同时通过分析训练样本集中各个样本作支持向量的次数的差异，来解释预报因子的物理意义及其在模型中的作用。

给定 C 和 g 的初值分别为 50 和 0.002，按一定的增量步长(分别为 10 和 0.002)和循环次数(10 次)自动改变 C 和 g 的值，从 489 个总样本中随机抽取 10%的样本作为检验样本，其余样本作为建模样本，共抽取 10 次，将上述 14 站各进行了 1 000 次交叉验证。交叉验证的结果表明 SVM 方法建立的预报模型对 14 站的降水都具有正的预报技巧，即模型有一定的预报能力；但对于不同的站点，预报的 TS 评分的差别还是比较大的，说明所选的 316 个因子并不是对每个站都适用。另外预报模型的稳定性还需在最优的 C 和 g 取值下进行多次的交叉验证才能得到。

8.2.2.3　SVM 模型中各因子的分析

为了分析各模型预报能力的差异，找出各因子的物理意义及其在模型中的作用，需对每个站点进行交叉建模试验，即在上述对每个站点的 1 000 次建模试验中，若有 900 次以上都被选为支持向量的样本（正、负样本均有，几乎各占一半，见表 8.4），求出这些样本中各因子的平均值。若在 1 000 次试验中从未被选为支持向量的样本（非支持向量样本，基本上是负样本），也从中求出各因子的平均值。这样做，意味着将样本划分出两类，前一类样本包含有明显的降水信息，后一类所包含的信息对降水不敏感。试图比较两类样本中各因子平均值的差异（类差异），从中挑出差异较大的预报因子，可以认为这些因子对降水预报的贡献较大，通过对这些因子的深入分析得到预报降水的有益启示。在样本包含的 316 个因子中，由于 4 个时次所选因子均相同，如果 4 个时次中有 3 个及其以上时次两类因子平均值的差别较大，则认为这个因子就是可选因子，这样实际最后只对 79 个因子进行分析。

表 8.4　1 000 次试验中支持向量的分布情况

站名	做支持向量样本在 900 次以上		未做支持向量样本	
	总个数	正样本个数	总个数	正样本个数
哈尔滨	69	35	38	0
大连	73	37	85	0
北京	63	35	84	0
郑州	62	24	182	0
武汉	49	25	180	1
南京	61	34	114	0
合肥	32	17	99	0
杭州	30	15	115	0
南昌	50	25	120	0
福州	55	27	133	0
厦门	43	26	154	0
广州	55	29	80	1
南宁	64	32	46	0
海口	54	28	50	0

所得结果大致可分为三种情况：

（1）在选取的 14 个代表站中，两类平均值的数值差别较大的因子有：850 hPa 涡度，1 000～600 hPa 累积水汽通量，1 000～600 hPa 累积水汽通量散度，1 000～850 hPa，400～200 hPa 和 700～500 hPa 的相对湿度的垂直累积量和螺旋度累积量，200～850 hPa 散度差和厚度平流差，700 hPa 3 时次累积平均垂直速度，1 000～600 hPa 3 时次累积水汽通量和 3 时次累积水汽通量散度等。

(2)在选取的 14 个代表站中,两类平均值的数值无差别或差别不大的有因子:1 000 hPa的 V 偏差风,1 000～300 hPa 各层湿位涡发展判据,850 和 700 hPa 的 K 指数,850 hPa沙氏指数,以及动力综合因子等。

(3)更多的因子是对某些站而言,两类平均值的数值有差别,对另一些站两类平均值的数值无差别(表略)。

在情况(1)所包含的因子中,各种累积量较多,也有高、低层之间的差值。说明单层或单时次的因子可能表现的热力动力过程不明确,但因子不同层次、不同时次的累积量或高低层之间的差值大,它们的物理意义明确,对降水的指示性较好,在选因子时应优先挑出。

在情况(2)所包含的因子中,以 K 指数、沙氏指数和湿位涡倾斜发展判据和动力综合因子为主,由于这些因子要么与对流关系密切,要么反映的动力过程太复杂,无法辨别它们的差异,在选降水因子时可以少选或不予考虑。

对于情况(3)反映出对大多数因子而言,在不同的站有不同的效果,要分不同的站来分析各个因子。但也有些因子如 U 偏差风的两类平均值的数值在南方差异比较大,而另外一些因子如 850 hPa V 偏差风两类平均值的数值在北方的差异大。因此,选因子时可有所侧重。

表 8.5 给出了 14 个站中两类平均值数值差异大的因子个数分布。从表中可以看出,除杭州、海口外,其他站两类平均值的数值差异大的因子个数都在 110～140 之间。为了找出这两站两类平均值数值差异大的因子个数少的原因,这里分析了1 000 次试验中 900 次以上做支持向量的样本正样本(≥10 mm)(表 8.4),发现尽管杭州和海口的正样本做支持向量的个数分别为 15 和 28,但强降水(≥50 mm)的支持向量个数(表略)分别为 2 和 0(不是最优 SVM)。由于强降水少,对应的物理量就可能与无降水时差异不大;同时 T213 的预报结果也存在预报误差,因此会造成这两个站两类平均值数值差异较小。反之,如果做支持向量的正样本中包含的强降水样本多,则两类平均值的数值差异就会较大。

表 8.5 14 站中两类平均值数值差异大的因子个数分布

站名	哈尔滨	大连	北京	郑州	武汉	南京	合肥	杭州	南昌	福州	厦门	广州	南宁	海口
个数	108	120	116	120	136	120	140	68	140	144	140	120	112	56

8.2.2.4 实例应用效果分析

分别用初选的 316 个预报因子、去掉 8.2.2.3 中情况(2)后的 264 个因子,以及表 8.5 中各站对应的因子个数分别建立 14 个站的降水预报模型,建模样本长度均为 489。然后用这些模型来试报 2006 年 4—8 月逐日的降水,表 8.6 给出了部分站的预报结果。

表 8.6　实例预报检验结果

站名	建模因子数	正样本TS评分	正样本正确次数	正样本空报次数	正样本漏报次数	负样本正确次数	正样本个数（比例）	分类准确率
哈尔滨	316	0.273	12	7	25	86	37(0.285)	0.754
	264	0.367	18	12	19	81	37(0.285)	0.762
	108	0.380	19	13	18	80	37(0.285)	0.762
大连	316	0.277	13	20	14	83	27(0.207)	0.738
	264	0.283	13	19	14	84	27(0.207)	0.746
	120	0.370	10	0	17	103	27(0.207)	0.869
北京	316	0.333	13	10	16	91	29(0.223)	0.800
	264	0.333	14	13	15	88	29(0.223)	0.785
	116	0.351	13	8	16	93	29(0.223)	0.815
福州	316	0.250	9	5	22	94	31(0.238)	0.792
	264	0.229	8	4	23	95	31(0.238)	0.792
	144	0.324	11	3	20	96	31(0.238)	0.823
广州	316	0.300	15	13	22	80	37(0.285)	0.731
	264	0.302	19	26	18	67	37(0.285)	0.662
	120	0.321	17	16	20	77	37(0.285)	0.723
南宁	316	0.118	4	12	18	96	22 (0.169)	0.769
	264	0.152	5	11	17	97	22 (0.169)	0.785
	112	0.310	9	7	13	101	22 (0.169)	0.846

从表 8.6 中可以看出，去掉一些因子后建立的预报模型对于哈尔滨、大连和华南的福州、广州、南宁的降水预报准确率提高比较明显。而用同样方法建立的模型对我国中部（武汉、南京、合肥、南昌）降水预报准确率提高不大（图略），或许是因为这些地区的降水影响系统更为复杂，去掉一些因子后，可能会漏掉一些有用的信息，所以对这些地区的降水预报还有待于进一步研究。

8.2.2.5　小结

通过基于 SVM 方法的参数和模型寻优功能进行了多次交叉验证试验，分析了各因子的意义。同时，用不同因子建立的降水模型来做实例预报，实例试报的结果表明挑选后的因子建立的预报模型能提高东北和华南降水的预报准确率；而对我国中部地区降水预报准确率则没有明显改进。

8.2.3　计算样本的贴近度

8.2.3.1　资料选取及处理

在全国选取 72 个气象站的日降水量作为预报对象，并读取这 72 个站 2003—2005 年 4—9 月及 2006 年 4 月 1 日—7 月 31 日逐日 08—08 时的观测降水量。选取了对应段内 T213 数值预报产品作为主要的预报因子，将 T213 的基本要素及其通过

动力诊断得出的反映降水的扩充物理量，用双线形插值的方法插值到对应的 72 个站点上，建立起所需要的站点因子库；再通过计算相关系数，在不同层次、不同时次的因子中选出一批与实况降水量相关系数较大的因子，然后按相关系数由大到小的顺序排列，选取 100 个左右的预报因子，这些因子中包含从 0，12，24 到 48 小时的预报时效因子。

这样就形成了 72 个站 2003—2005 年 4—9 月、2006 年 4 月 1 日—7 月 31 日共 600 个左右的历史学习训练样本集。

另外，以 2006 年 8 月 1 日—9 月 30 日共 60 个样本作为独立检验样本集。

8.2.3.2 降水预报模型的建立及模型的检验

这里用 SVM 两类分类方法，在建模之前先对预报对象进行分类。

由于西北地区降水较少，因此将西北 9 个站（乌鲁木齐、克拉玛依、酒泉、民勤、兰州、呼和浩特、盐池、太原、银川）日降水量≥1 mm 标定为正样本（+1 类）、日降水量＜1 mm 标定为负样本（−1 类）；其余 63 个站日降水量≥10 mm 标定为正样本（+1 类）、日降水量＜10 mm 标定为负样本（−1 类）。再对每个站全部样本的每一个因子分别做归一化处理，使每个因子的数据在[0，1]之间，这样避免了预报因子之间量级的差异。

选取最常用的径向基核函数，通过调整核参数 g 和惩罚系数 C 的值，进行大量随机交叉验证，分析比较所建模型 TS 评分的高低，从而选择出最优模型对应的参数 g 和 C。

从全部 600 个样本中随机抽取 10%的样本作为检验样本，其余样本作为建模样本，C 和 g 给定的初值分别为 100 和 0.005，再按一定的步长递增（分别为 5 和 0.005），对每个站进行 1 000 次随机交叉验证，从中选出最高 TS 评分值对应的参数 C 和 g。然后以 2003—2005 年 4—9 月、2006 年 4 月 1 日—7 月 31 日共 600 个左右的训练样本集分别建立各站的降水预报模型，再用 60 个独立的样本来作做预报试验，检验各预报模型的预报能力，即预报 2006 年 8 月 1 日—9 月 30 日间逐日的降水，部分站预报评分结果见表 8.7；T213 模式直接输出的降水预报（Direct Model Output，DMO）评分检验也列于表 8.7。

表 8.7 预报试验期间部分站 SVM 预报和 T213 DMO 的 TS 评分对比

站 名	SVM 预报	T213DMO
民勤	0.308	0.207
兰州	0.462	0.286
太原	0.444	0.394
哈尔滨	0.500	0.333
漠河	0.571	0.333
呼和浩特	0.375	0.320

续表

站 名	SVM 预报	T213DMO
平凉	0.367	0.250
本溪	0.400	0.364
北京	0.500	0.067
济南	0.500	0.286
泰山	0.440	0.308
昌都	0.333	0.000
略阳	0.333	0.077
汉中	0.545	0.444
安康	0.600	0.364
遵义	0.333	0.000
安庆	0.500	0.000
河池	0.429	0.400
漳州	0.556	0.217
厦门	0.375	0.176
广州	0.316	0.286
海口	0.500	0.294
三亚	0.333	0.231

独立样本检验的结果(表 8.7)表明,SVM 建立的模型对降水有较好地预报能力,且比 T213 模式直接输出的降水预报的 TS 评分高,这些站的降水预报模型可以投入业务试运行。

8.2.3.3　贴近度分析及模型的改进

利用贴近度计算公式[式(7.5)],分别计算了 72 个站的贴近度值。由于 72 个站中正、负样本数不同,组成样本的因子也各不相同,所以经过计算得到的样本间的贴近度也不同(表略)。

这里重点对比分析 SVM 模型在独立检验时 TS 评分较低、同类样本之间的样本贴近度小、正负样本间的贴近度大的站,从而实现预报因子的筛选。

SVM 方法要求同类样本的贴近度尽可能大,不同类样本的贴近度尽可能小。以下是对 20 个站中同类样本贴近度小、不同类样本贴近度较大的样本中各因子分析的结果,这些因子在样本中应该去除。它们大致可以分为下面 6 类:

(1) 一些较复杂的热力、动力因子:如螺旋度、$\boldsymbol{Q}$ 矢量、动力综合因子、锋生函数、位涡及湿位涡有关的量、风的垂直切变、风向、K 指数、SI 指数、K_y 指数、温度的指数、比湿的指数、U 和 V 的指数等;

(2) 层次高的因子:如 200 hPa 的相当位温、假相当位温的垂直切变、50 hPa 的 U 和 V,150 hPa 的温度及其梯度;

(3) 近地层因子(特别是热带地区):如 1.5 m 的温度、湿度,海平面气压、地面气压及变压,10 m 的风,1 000 hPa 的变温、变高等;

(4) 高低层差值和累积量:如 200 和 850 hPa 的高度差,400,300 和 200 hPa 垂直速度的累积量,700,600 和 500 hPa 温度的累计量,地面到 600 hPa 水汽通量及水汽通量散度等;

(5) 同一层次、同一要素、同一性质的因子太多:如 700 hPa 温度的平方和立方及 e 指数,相对湿度、比湿的平方和立方及 e 指数被同时选作因子等;

(6)海拔高度以下的因子:如高山站海平面气压、地面气压、850 hPa 的风速和风向等;

用贴近度计算分析后,去掉一些因子,各站筛选后的预报因子个数各不相同(表略),用筛选后的因子重新组成新的训练建模样本集,再进行 1 000 次随机交叉验证,得到最优的参数 C 和 g,从而得到改进后的预报模型。改进后的预报模型对 2006 年 8 月 1 日—9 月 30 日逐日降水的预报效果的检验见表 8.8,表 8.8 也给出了同一时段内 SVM 模型改进前降水预报、T213 模式直接输出的降水预报(DMO)结果的检验。

表 8.8 20 个站 SVM 模型改进前后及 T213DMO 的 TS 评分对比

站 名	SVM 模型改进后	SVM 模型改进前	T213 DMO
乌鲁木齐	0.222	0.167	0.167
银川	0.625	0.200	0.250
石家庄	0.714	0.250	0.500
长春	0.375	0.200	0.333
拉萨	0.250	0.000	0.000
贵阳	0.333	0.222	0.182
井冈山	0.200	0.143	0.154
南京	0.250	0.200	0.111
福州	0.400	0.125	0.227
九仙山	0.357	0.214	0.167
南宁	0.286	0.200	0.273
北海	0.267	0.154	0.250
汕头	0.333	0.214	0.250
桂林	0.286	0.143	0.143
酒泉	0.200	0.200	0.143
郑州	0.143	0.125	0.250
徐州	0.200	0.000	0.111
杭州	0.167	0.000	0.167
武夷山	0.143	0.143	0.133
湛江	0.167	0.090	0.111

由表 8.8 可知,改进后模型的 TS 评分基本上比改进前的 TS 评分和 T213 模式直接输出的降水预报 TS 评分高,表明改进后的模型对降水有一定的预报能力,但仍

有少数站的预报效果较差。

8.2.3.4　实时预报效果的检验

2007 年 8 月 1 日开始，SVM 方法建立的降水预报模型在国家气象中心投入业务试运行，对于预报试验期间（即 2006 年 8 月 1 日—9 月 30 日逐日降水预报）TS 评分＞0.25 的站，用原来的预报因子（100 个左右）组成的样本集建模；对于预报试验期间 TS 评分≤0.25 的站，则用贴近度分析后筛选的因子组成的样本集建模。为了增强所建模型的稳定性，实时业务运行时，增加了建模样本的长度，这里使用了 2003—2005 年 4—9 月、2006 年 4 月 1 日—7 月 31 日共 600 个左右的原训练样本集和 2006 年 8 月 1 日—9 月 30 日独立检验的 60 个样本，即建模训练样本的长度为 2003—2006 年 4—9 月 660 个样本的历史资料。共有 72 个站的降水预报模型在 2007 年 7 月 31 日—9 月 30 日进行了实时业务试运行，部分站的预报检验结果见表 8.9，表 8.9 也给出了同一时段内 T213 模式的直接降水预报（DMO）结果的检验。

表 8.9　实时预报中部分站 SVM/T213 DMO 的检验结果

站名	正样本 TS 评分	正样本 正确次数	正样本 空报次数	正样本 漏报次数	负样本 正确次数	正样本 个数	全样本 准确率
乌鲁木齐	0.333/0.286	3/6	2/14	4/1	52/40	7/7	0.902/0.754
兰州	0.429/0.293	6/12	2/29	6/0	47/20	12/12	0.869/0.525
盐池	0.412/0.407	7/11	4/14	6/2	44/34	13/13	0.836/0.738
太原	0.412/0.381	14/8	18/5	2/8	27/40	16/16	0.672/0.787
长春	0.600/0.375	3/3	2/5	0/0	56/53	3/3	0.967/0.918
沈阳	0.308/0.308	4/4	7/7	2/2	48/48	6/6	0.852/0.852
北京	0.250/0.125	1/1	0/4	3/3	57/53	4/4	0.951/0.885
大连	0.364/0.353	4/6	2/8	5/3	50/44	9/9	0.885/0.820
泰山	0.421/0.211	8/4	9/9	2/6	42/42	10/10	0.820/0.754
宜昌	0.231/0.190	3/4	3/11	7/6	48/40	10/10	0.836/0.721
重庆	0.556/0.400	5/4	4/5	0/1	52/51	5/5	0.934/0.902
郴州	0.308/0.231	4/3	6/6	3/4	48/48	7/7	0.852/0.836
徐州	0.250/0.190	3/4	3/12	6/12	49/40	9/9	0.852/0.721
庐山	0.429/0.333	6/6	1/5	7/7	47/43	13/13	0.869/0.803
福州	0.200/0.192	3/5	8/19	4/2	46/35	7/7	0.803/0.656
九仙山	0.500/0.515	13/17	5/12	8/4	35/28	21/21	0.787/0.738
河池	0.460/0.300	7/6	7/9	4/5	46/41	11/11	0.869/0.770
漳州	0.273/0.235	6/8	8/20	8/6	39/27	14/14	0.738/0.574
厦门	0.313/0.258	5/8	8/23	3/0	45/30	8/8	0.820/0.623
汕头	0.471/0.385	8/10	4/13	5/3	44/35	13/13	0.852/0.738
北海	0.308/0.273	4/6	1/10	8/6	48/39	12/12	0.852/0.738
三亚	0.286/0.200	4/6	3/19	7/5	47/31	11/11	0.836/0.607
珊瑚岛	0.357/0.304	5/7	4/13	5/3	47/38	10/11	0.852/0.738

实时业务试运行的结果(表 8.9)也表明 SVM 建立的模型对降水具有预报能力，其 TS 评分和预报准确率也基本上高于 T213 模式直接输出的降水预报。相对而言，SVM 方法的漏报较多，而 T213 模式的空报较多。

8.2.3.5 小结

由上述的分析可知，通过样本间贴近度的分析，去掉一些因子，实现因子的筛选和预报模型的改进，从而提高了模型的预报能力。

8.3 SVM 方法用于短期气候预测

目前 SVM 方法在短期气候预测中的应用主要表现在雾气候趋势预测、冬季温度趋势预测、汛期降水预测等方面。这里仅给出了该方法在雾气候趋势预测中的实例(常军等 2008)。

8.3.1 资料和方法

平均值取 1971—2000 年 30 年平均，首先对郑州 1951—2005 年冬半年(10 月 — 次年 3 月)雾日数气候特征进行分析，选择气候预测中常用的 74 项环流特征量(资料来自国家气候中心)作为备选因子，经过对预测因子的筛选，最终选用 4 个预测因子，用 1951/1952—2001/2002 年共 51 年作为建模样本，2002/2003—2005/2006 年 4 年作为独立样本进行试报，分别采用 SVM 回归方法建立预测模型，对冬半年大雾日数进行预测。

8.3.2 预测因子的构建

由于冬半年(10 月—次年 3 月)的预测一般是在 9 月制作，因此选择前期 1—8 月的资料作为备选因子较好(74 个×8 = 592 个)。充分考虑因子的代表性后，选取 4 个预测因子用于建模。这 4 个预测因子分别是 5 月北非大西洋北美副高北界、3 月冷空气次数、5 月黑潮区(125°～160°E，25°～35°N)海温指数、2 月 5 区(100°W～40°E，45°～65°N)阻塞高压面积指数，它们与冬半年雾日数的相关系数分别是 −0.58，−0.4，−0.41、−0.43，均通过了 $\alpha = 0.01$ 的显著性检验。为保证尽量多的样本参与建模，在现有数据集(55 年观测资料)中，用 51 个样本作为验证数据文件，留出 4 个样本做预报，以径向基函数作为 SVM 方法中的核函数建立推理试验模型。通过逐渐改变参数的取值，最后综合所有样本验证的结果，选取最优参数和最优模型。具体回报和预测结果见图 8.3。

8.3.3 雾的预测评定方法及独立样本检验

我国目前气候预测业务只评定降水和温度(降水和温度预测的分级有 6 级评分制和 4 级评分制)，还没有进行雾、大风、沙尘、高低温日数等的评定。为与国际短期气候预测评分办法接轨，提高“气候趋势”预测能力，注意与目前短期气候预测水平及

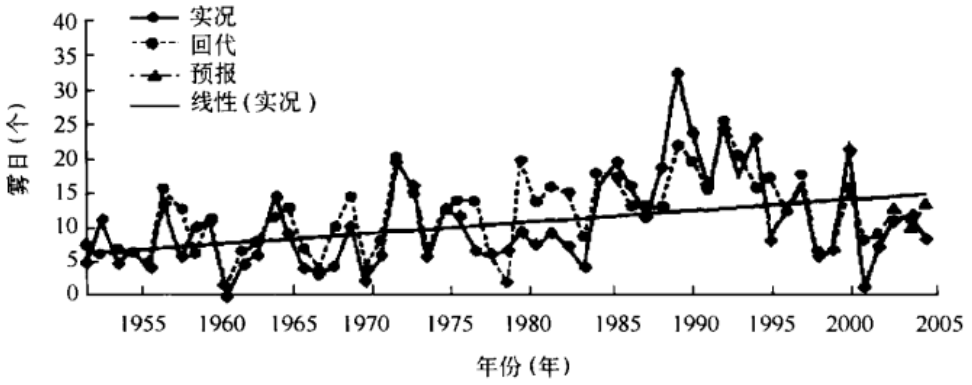


图 8.3　SVM 回归模型预报结果与实况结果对比

业务现状衔接，我们参考现行气候预测业务评分标准，考虑雾日基数比降水小，对雾日数采取 4 级评定制，4 级预测用语和各等级划分标准见表 8.10，评定结果见表 8.11。因郑州冬半年雾日数平均为 16 天，因此按表 1 可划分为：偏少为≤11 天，略少为＞ 11 天且＜ 16 天，略多为≥16 天且＜ 21 天，偏多为≥21 天。

表 8.10　4 级预测用语及各等级划分标准

预测用语	偏少	略少	略多	偏多
距平百分比(%)	≤－30	＞－30 且＜0	≥0 且＜30	≥30

表 8.11　4 级评分制得分统计

项目	预测偏少	预测略少	预测略多	预测偏多
实况偏少	100	100	0	0
实况略少	100	100	100	0
实况略多	0	100	100	100
实况偏多	0	0	100	100

按评分标准评定，SVM 回报正确率为 82.14%（表略）。经独立样本试报，2002/2003—2005/2006 年(SUM 方法平均)预测准确率(TS)为 75% ，前 3 年预测正确，第 4 年报错(具体情况见表 8.12)。

表 8.12　SVM 方法在 2002/2003—2005 /2006 年试报评定结果

预报时段	实况	预测	误差	评定(%)
2002/2003	9	11	2	100
2003/2004	13	15	2	100
2004/2005	14	12	－2	100
2005/2006	10	16	6	0

注：实况、预测、误差对应列数字，均为雾日(个)

8.3.4 小结

SVM 回归方法对郑州冬半年雾的气候预测试应用效果均较好，回报正确率在80%以上，独立样本试报准确率都达到了 75% ，预测绝对误差平均只有 310 天，该方法可在业务中试应用。

8.4 SVM 与 ANN 方法预报效果的比较

本节以天空云量预报的建模为例，将 SVM 与人工神经网络(Artrficial Neural Network, ANN)方法在模型的预报能力、算法的复杂性、模型的鲁棒性及泛化能力等方面进行了对比分析，并将其用于实际预报检验。

8.4.1 所建模型的预报能力及稳定性分析

8.4.1.1 SVM 预报模型

利用熊秋芬等(2007)中的 2001 年 5 月 1 日—2004 年 12 月 31 日共 1 340 个样本的历史资料作为训练学习建模分析数据，进行随机交叉验证。这里仅列出了选取参数 $C=100$ 且 $g=0.01$(日平均总云量<4 成)，$C=100$ 且 $g=0.032$(日平均总云量≥8 成)时分别进行的 50 次交叉验证的结果，也就是对于日平均总云量<4 成(占总样本的 21.27%)和日平均总云量≥8 成(占总样本的 48.88%)的两种情况，分别抽取 10%，20%和 30%的样本作为检验数据，按正样本的 TS 评分寻优标准各进行了 50 次随机试验，共获得 300 个试验模型的结果数据，其间每次计算时间不超过 10 秒。表 8.13 给出了这些试验中平均 TS 评分的统计结果。由于日平均总云量≥8 成的分类中，正、负样本的比例差不多(接近 50%)，同样抽取 10%，20%和 30%的样本作为检验数据，按全样本的分类准确率寻优标准进行了随机试验，共包括 150 次试验模型的结果数据，所以表 8.14 给出了全样本的预报准确率的平均统计结果。

从表 8.13 中可以看出，虽然单次检验的 TS 评分有高低起伏，但 SVM 方法对日平均总云量<4 成的 TS 评分的平均值分别为 0.421 8，0.410 8 和 0.405 7，比实际样本的 21.27%提高了近 20%，正预报技巧显著；日平均总云量≥8 成的 TS 评分的平均值分别为 0.669 0，0.662 4 和 0.649 7，比实际样本的 48.88%提高了约 17%，也表现出明显的正预报技巧；全样本的预报准确率在 80%左右(表 8.14)。这些都表明 SVM 方法建立的模型有较好地预报能力且是稳定的。

另外随着抽取的检验样本数的增加，建模训练样本数的减少，两种分类情况下 TS 平均值(准确率)的变化不大，并没有明显减少，而对应的 TS 评分(准确率)的平均绝对偏差却在明显地减少(表 8.13 和 8.14)，也表明该方法建立的模型是稳定的，同时也说明 SVM 方法建立的模型具有推广应用能力。

表 8.13　SVM 交叉验证结果统计

交叉验证样本比例		10%	20%	30%
日平均总云量<4 成	平均 TS 评分	0.421 8	0.410 8	0.405 7
	TS 平均绝对偏差	0.082 1	0.040 2	0.031 6
日平均总云量≥8 成	平均 TS 评分	0.669 0	0.662 4	0.649 7
	TS 平均绝对偏差	0.034 2	0.022 7	0.020 7

注：参数 $C=100, g=0.032$

表 8.14　日均总云量≥8 成的 SVM 交叉验证准确率统计结果

交叉验证样本比例	10%	20%	30%
平均准确率/%	79.66	80.26	79.59
准确率平均绝对偏差	0.025 5	0.018 4	0.013 5

注：参数 $C=100, g=0.032$

8.4.1.2　ANN 预报模型

为了便于与 SVM 方法的结果进行对比，分别从 SVM 的日平均总云量<4 成和日平均总云量≥8 成两种情况下按 10%，20%和 30%的比例随机抽取的 50 次检验样本号和建模样本号中，各选出 30 次分别作为 ANN 方法的检验样本号和建模样本号（本应选择同样的次数，但 ANN 建模耗时太多，只好从简），来建立 ANN 的建模样本集和检验样本集，进行对比试验。这相当于对日平均总云量<4 成和日平均总云量≥8 成两种情况，各进行了 30 次随机试验。同样为了避免各预报因子之间量级的差异，在用 ANN 方法建模之前，对全部样本的每一个预报对象和预报因子分别做了归一化处理，使每个数据在[0,1]之间。计算时迭代次数为 50 000 次，隐含层节点数为 10，每次计算时间约需 15 分钟（900 秒）。共获得 180 个试验模型的结果数据，平均 TS 评分的统计结果见表 8.15。由于同样的原因表 8.16 也列出了全样本 90 次实验模型预报准确率的统计结果。

表 8.15　ANN 交叉验证的结果简表

交叉验证样本比例		10%	20%	30%
日平均总云量<4 成	平均 TS 评分	0.332 5	0.371 2	0.349 8
	TS 平均绝对偏差	0.051 1	0.038 4	0.031 2
日平均总云量≥8 成	平均 TS 评分	0.594 0	0.584 0	0.565 9
	TS 平均绝对偏差	0.028 6	0.029 7	0.017 9

表 8.15 表明，ANN 方法对日平均总云量<4 成的 TS 评分的平均值分别为 0.332 5，0.371 2 和 0.349 8，比实际样本的 21.27%提高了近 12%～15%，有明显的正预报技巧；日平均总云量≥8 成的 TS 评分的平均值分别为 0.594 0，0.584 0 和 0.565 9，比实际样本的 48.88%提高了近 10%，也具有正预报技巧；全样本的预报准确率分别为 72.43%，72.26%和 70.49%（表 8.16），表明 ANN 方法建立的模型具有

良好的预报能力且是稳定的。TS评分(准确率)的平均绝对偏差逐渐减小的趋势也说明ANN方法建立的模型是稳定的,具有一定的推广应用能力。

表 8.16 日均总云量≥8成的ANN交叉验证准确率统计结果

交叉验证样本比例	10%	20%	30%
平均准确率(%)	72.43	72.26	70.49
准确率平均绝对偏差	0.021 0	0.018 6	0.018 3

与SVM方法的评分结果相比,虽然ANN有少数检验的评分偏高(表略),但总体上比SVM方法的TS评分(准确率)要低5%～9%。

8.4.2 实例应用效果分析

以2001年5月1日—2004年12月31日的1 340个样本资料作为训练建模样本,分别用SVM和ANN方法建立预报模型,再将2005年1月1日—5月31日逐日20时武汉市地面、高空观测值和ECMWF 24小时预报场等资料组成的81个预报因子分别做归一化处理后(实际只有141天资料),分别输入预报模型来预报第二天武汉市的日平均总云量。用SVM方法进行计算时,对于日平均总云量<4成的情况,核参数 $g=0.01$,$C=100$;对于日平均总云量≥8成的情况,核参数 $g=0.039$,$C=100$。用ANN方法计算时迭代次数仍为50 000次,隐含层节点数取10,计算结果见表8.17。

表8.17表明,SVM和ANN方法对天空云量都有较好的预报能力,它们对正样本预报正确的次数和漏报的次数基本相同,但是ANN对正样本的空报次数稍多,对负样本预报的正确次数略少,所以SVM方法的预报效果比ANN方法好,尤其在正样本少(日平均总云量<4成)的情况下,SVM方法优势更明显。

表 8.17 SVM/ANN实例预报结果

分类	日平均总云量<4	日平均总云量≥8
正样本TS评分	0.464/0.406	0.716/0.705
正样本正确次数	13/13	73/74
正样本空报次数	6/10	22/25
正样本漏报次数	9/9	7/6
负样本正确次数	113/109	39/36
正样本个数及占全部样本的比例(%)	22(15.6%)	80(56.73%)
全样本准确率(%)	89.4/86.5	79.4/78.0

8.4.3 小结

通过对SVM和ANN方法建立的天空云量模型及预报结果的比较分析,得到如下结论:

(1) SVM 和 ANN 模型都是比较稳定的，也表明这两种模型对天空云量有较好地预报能力和推广应用能力。

(2) 无论是对历史资料的交叉验证，还是用于预报实例，SVM 方法的评分结果均优于 ANN 方法，表明 SVM 比 ANN 方法更适用于天空云量的预报；并且 SVM 方法在计算速度方面具有 ANN 方法无法比拟的优越性。

8.5　本章小结

第 8 章仅介绍了支持向量机方法在短期温度和降水预报、预报因子选取及雾的短期气候预测中应用的部分实例。但由于其应用于气象预报的时间不长，特别是要将该方法投入日常预报业务使用时，在如何采取动力和统计相结合的释用技术来选取预报因子、改进预报预测模型，逐步提高灾害性天气气候的预报预测能力等方面，仍有许多工作有待于进一步完善和提高。

自 2004 年 SVM 的分类和回归方法被首次引入气象预报预测领域以来，该方法受到广大气象业务和科研工作者的关注，很快在晴雨预报、降水量预报、暴雨预报、面雨量预报、雷雨冰雹预报、台风路径预报、温度预报、天空云量预报、能见度和雾的预报、大气污染预报、水文气象预报、汛期旱涝预预测、沙尘暴预测、场预测、短期气候预测，以及预报因子分析等广泛的领域得到实际应用。本书附录 2 列出了截至 2009 年底在相关学术期刊刊载的相关应用论文 44 篇(不完全统计)，它们发表于《气象学报》、《应用气象学报》、《热带气象学报》、《气象科技》、《气象》、《干旱气象》和《暴雨灾害》等专业刊物。

随着时代和技术的推进，新的气象业务和服务需求给我们提出了更多的要求，当我们面对新课题只有历史资料而没有可用的预报模型时，运用统计学习方法来解决问题是最佳选择。我们殷切期待着 SVM 及核方法在基础研究和业务应用上不断提高水平，年年达到新的境界。

我们也热切欢迎广大 CMSVM 的用户根据实际情况，对 CMSVM 2.0 版本提出改进要求和建议，促使我们的工作不断深入。

参考文献

常军，李素萍，李祯，等. 2008. CAR 和 SVM 方法在郑州冬半年大雾气候趋势预测中的试用. 气象与环境科学，**31**(1)：17-19.

陈永义. 2005. 支持向量机方法与模糊系统. 模糊系统与数学，**19**(1)：1-11.

陈永义，俞小鼎，高学浩，等. 2004. 处理非线性分类和回归问题的一种新方法(I)——支持向量机方法简介. 应用气象学报，**15**(3)：345-354.

Courant R，Hilbert D. 1958. 数学物理方法(卷 I). 钱敏，郭敦仁译. 北京：科学出版社.

Cristianini N，Shawa-Taylor J. 2004. 支持向量机导论. 李国正，王猛，曾华军译. 北京：电子工业出版社.

冯·诺伊曼. 2010. 计算机和人脑. 甘子玉译. 北京：北京大学出版社.

冯汉中，陈永义. 2004. 处理非线性分类和回归问题的一种新方法(II)——支持向量机方法在天气预报中的应用. 应用气象学报，**15**(3)：355-365.

Hastie T，Tibshirani R，Friedman J. 2004. 统计学习基础：数据挖掘、推理与预测. 范明，柴玉梅，昝红英，等译. 北京：电子工业出版社.

黄嘉佑. 2000. 气象统计分析与预报方法(第二版). 北京：气象出版社.

Vapnik V N. 2000. 统计学习理论的本质. 张学工译. 北京：清华大学出版社.

Vapnik V N. 2004. 统计学习理论. 许建华，张学工译. 北京：电子工业出版社.

王建生，熊秋芬. 2007. 支持向量机方法在单站降水预报中的应用探讨. 暴雨灾害，**26**(2)：159-162.

熊秋芬，高杰，刘还珠，等. 2007a. SVM 方法预报降水的因子物理意义分析. //中国气象局预测减灾司. 天气预报技术文集(2007). 北京：气象出版社.

熊秋芬，顾永刚，王丽. 2007b. 支持向量机分类方法在天空云量预报中的应用. 气象，**33**(5)：20-26.

熊秋芬，胡江林，陈永义. 2007c. 天空云量预报及支持向量机和神经网络方法比较研究. 热带气象学报，**23**(3)：255-260.

熊秋芬，曾晓青. 2008. SVM 方法在降水预报中的应用及改进. 气象，**34**(12)：90-95.

Burges C J C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*，**2**(2)：127-167.

Chen Y X，Wang J Z. 2003a. Kernel machines and additive fuzzy systems：Classification and function approximation. *IEEE International Conference on Fuzzy Systems*，**2**：789-795.

Chen Y X，Wang J Z. 2003b. Support vector learning for fuzzy Rule-Based classification systems. *IEEE Transactions on Fuzzy Systems*，**11**(6)：716-728.

Chen Y Y，Feng H Z. 2003. A kernel method for fuzzy systems modeling and approximate reasoning. NAFIPS，Chicago.

Joachims T. 1998. Text categorization with support vector machines: Learing with many relevant features. *Machine Learning*, **1398**(26): 137-142.

Joachims T. 1999. Making large-scale SVM learning practical. Advances in Kernel Methods. MIT Press, Cambridge.

Mangasarian O L. 2000. Generalized support vector machines. // Peter J. Bartlett, Bernhard Schölkopf, Dale Schuurmans, etc, eds. Advances in Large Margin Classifiers, MIT Press, Cambridge.

Mercer J. 1909. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society* (A), **209**: 415-446.

Meyer D, Leisch F, Hornik K. 2002. Benchmarking Support Vector Machines, Report No. 78. Vienna University of Economics and Business Administration, Austria

Muller K R, Mika S, Ratsch G, et al. 2001. An introduction to Kernel-Based learning algorithms. *IEEE Transactions on Neural Networks*, **12**(2): 181-202.

Schölkopf B, Burges C J C, Smola A J. 1999a. Advances in Kernel Methods—Support Vector Learning. MIT Press, Cambridge.

Schölkopf B, Williamson R, Smola A J, et al. 1999b. Support vector method for noveity detection. // Solla S A, Leen T K, Müller K R, eds. *The Neural Information Processing Systems*, **12**: 582-588.

Smola A J, Schölkopf B. 1998a. A tutorial on support vector regression. NeuroCOLT2 Technical Report Series, NC2-TR-1998-030.

Smola A J, Schölkopf B. 1998b. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, **22**: 211-231.

Vidyasagar M. 1997. A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems. Springer-Verlag, New York.

附录　支持向量机方法在天气预报中应用论文索引

（2004—2009 年不完全统计）

常军，李素萍，李祯，等. 2008. CAR 和 SVM 方法在郑州冬半年大雾气候趋势预测中的试用. 气象与环境科学，**31**(1)：17-19.

常军，李祯，朱业玉，等. 2005. 基于支持向量机方法（SVM）的冬季温度预测. 气象科技，**33**（增刊）：101-104.

常涛. 2006. 支持向量机在大气污染预报中的应用研究. 气象，**32**(12)：61-65.

车怀敏. 2005. 用支持向量机机方法作德阳降水预报. 四川气象，**25**(2)：13-15.

陈淑琴，张阑廉，黄辉. 2007. 用支持向量机方法预报舟山沿海突发性冷空气大风. 海洋预报，**24**(1)：51-55.

陈晓艳，赵玉金，孙文英，等. 2006. 支持向量机方法作温度预报试验. 贵州气象，**30**(1)：31-33.

陈永义，俞小鼎，高学浩，等. 2004. 处理非线性分类和回归问题的一种新方法（I）——支持向量机方法简介. 应用气象学报，**15**(3)：345-354.

冯汉中，陈永义. 2004. 处理非线性分类和回归问题的一种新方法（II）——支持向量机方法在天气预报中的应用. 应用气象学报，**15**(3)：355-365.

冯汉中，陈永义. 2005. 支持向量机回归方法在实时业务预报中的应用. 气象，**31**(1)：41-44.

冯汉中，陈永义，成永勤，等. 2006. 双流机场低能见度天气预报方法研究. 应用气象学报，**17**(1)：94-99.

冯汉中，杨椒群，刘波. 2005. 支持向量机（SVM）机方法在气象预报中的个例试验. 四川气象，**25**(2)：9-12.

贺皓，罗慧. 2009. 基于支持向量机模式识别的大雾预报方法. 气象科技，**37**(2)：149-151.

黄奕铭. 2006. 支持向量机在雷雨天气预报中的应用. 广东气象，**28**(1)：22-24.

李才媛，韦惠红，邓红. 2008. SVM 方法在武汉市大雾预警预报中的应用. 暴雨灾害，**27**(3)：264-267.

李素萍，常军，朱业玉，等. 2006. 支持向量机方法在郑州冬季气温趋势预测中试应用. 河南气象，(1)：15-16.

李智才，马文瑞，李素敏，等. 2006. 支持向量机在短期气候预测中的应用. 气象，**32**(5)：57-61.

刘科峰，张韧，洪梅，等. 2009. 基于最小二乘支持向量机的副热带高压预测模型. 应用气象学报，**20**(3)：354-359.

刘科峰，张韧，徐海斌，等. 2007. 支持向量机与卡尔曼滤波集合的西太平洋副热带高压数值预报误差修正. 气象学报，**65**(3)：450-457.

刘科峰，张韧，于鹏，等. 2007. 基于小波分解和最小二乘支持向量机的西太平洋副高预测模型. 热带气象学报，**23**(5)：491-496.

路志英，张启孟，赵智超. 2006. 基于 SVM 的沙尘暴预测模型. 天津大学学报，**39**(9)：1 110-1 114.

吕庆平，罗坚，朱坤，等. 2009. 基于 SVM 的气候持续法在热带气旋路径预报中的应用试验. 海洋预报，**26**(1)：76-83.

罗伟，习华勇. 2008. 基于最小二乘支持向量机的降雨量预测. 人民长江，**39**(19)：29-31.

罗泽举，宋丽红，薛宇峰，等. 2006. 一类基于 SVM/RBF 的气象模型预测系统. 计算机工程，**32**(21)：31-32，93.

盛文斌，张瑞萍，陆琛莉. 2008. SVM 方法在热带气旋降水预报中的应用. 广东气象，**30**(3)：6-8.

孙田文，房春琴，燕东渭. 2006. 支持向量机回归方法在铜川降水量预报中的应用. 陕西气象，(6)：4-7.

滕卫平，俞善贤，胡波. 2008. SVM 回归法在汛期旱涝预测中的应用研究. 浙江大学学报(理学版)，**35**(3)：343-354.

滕卫平，俞善贤，胡波. 2008. 基于支持向量机的浙江汛期旱涝预测方法研究. 气象科技，**36**(2)：139-144.

王建生，熊秋芬. 2007. 支持向量机方法在单站降水预报中的应用探讨. 暴雨灾害，**26**(2)：159-162.

王小萍，谭季青，吴书成. 2006. 基于支持向量机的预报技巧的预报模型研究. 科技通报，**22**(6)：747-752.

韦惠红，李才媛，邓红，等. 2008. SVM 方法在武汉区域 5—7 月暴雨预报中的业务应用. 2008 年全国重大天气预报经验交流会文集. 北京：气象出版社.

韦惠红，李才媛，邓红，等. 2009. SVM 方法在武汉区域夏季暴雨预报业务中的应用. 气象科技，**37**(2)：145-148.

吴爱敏，郭江勇，张洪芬，等. 支持向量机在冰雹预报中的应用. 干旱气象，2005，**23**(4)：41-45.

谢五三，田红. 2007. 基于支持向量机的安徽汛期降水预测研究. 安徽气象(电子版)，04. www.ahqx.gov.cn/ahqxoa/office/ahqx/2007_04/.

熊秋芬，高杰，刘还珠，等. 2007. SVM 方法预报降水的因子物理意义分析. 2007 年全国重大天气预报经验交流会文集. 北京：气象出版社.

熊秋芬，顾永刚，王丽. 2007. 支持向量机分类方法在天空云量预报中的应用. 气象，**33**(5)：20-26.

熊秋芬，胡江林，陈永义. 2007. 天空云量预报及支持向量机和神经网络方法比较研究. 热带气象学报，**23**(3)：255-260.

熊秋芬，曾晓青. 2008. SVM 方法在降水预报中的应用及改进. 气象，**34**(12)：90-95.

燕东渭，孙田文，杨艳等. 2007. 支持向量数据描述在西北暴雨预报中的应用试验. 应用气象学报，**18**(5)：676-682.

燕东渭，杨艳，孙田文，等. 2006. SVM 和 ANN 方法在铜川盛夏暴雨预报中的对比试验. 陕西气象，(3)：15-17.

杨淑群，芮景析，冯汉中. 2006. 支持向量机(SVM)方法在降水分类预测中的应用. 西南农业大学学报(自然科学版)，**28**(2)：252-257.

杨文发，周新春，段红. 2007. 三峡水库中期水文气象耦合预报应用试验及探讨. 水文，**27**(3)：39-42.

张春花，林建兴，郝丽清. 2007. 支持向量机方法在海口降水预报中的应用研究. 气象研究与应用，

28(增刊Ⅰ):7-9.

张礼平,陈永义. 2006. 支持向量机(SVM)及其在场预测中的应用. 热带气象学报,**22**(3):278-282.

赵国令,肖科丽. 2004. 支持向量机方法在天气预报中的应用. 陕西气象,(6):1-3.